POLITECNICO DI TORINO

III Facoltà di Ingegneria dell'Informazione
Corso di Laurea in Ingegneria delle Telecomunicazioni

Tesi di Laurea Magistrale

MISURE DEL TRAFFICO INTERNET: STIMA DELLA MATRICE DI TRAFFICO IN RETI REALI

Relatore:
prof. Marco Mellia

Candidato:
Marta Calmet

Gennaio 2009

Ringraziamenti

Vorrei ringraziare ad amici e colleghi per avermi supportato in questo periodo della mia formazione accademica. Ci sono stati giorni difficili e senza il supporto di tante persone sarebbe stato più duro andare avanti. È per ciò che vorrei ringraziare professori, amici e parenti che sono stati appoggiandomi giorno dopo giorno per la conclusione di questo duro percorso fino ad oggi. Volevo ringrazziare i colleghi ed amici conosciuti in Motorola, per averci stato sempre, specialmente in questi ultimi giorni in cui sono stata difficile da sopportare. E in fine, volevo ringraziare a Robert e il prof. Mellia, che sono stati guide fondamentali nell'elaborazione di questa tesi.

A tutti voi, *Grazie!!*

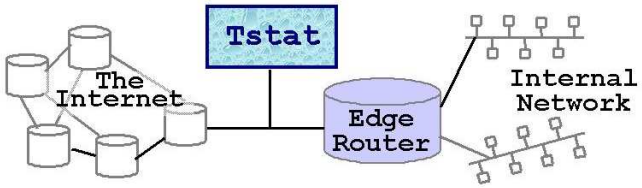
MISURE DEL TRAFFICO INTERNET: STIMA DELLA MATRICE DI TRAFFICO IN RETI REALI

Candidato: Marta Calmet
Relatore: Prof. Marco Mellia

Questa tesi nasce dalla necessità di analizzare una gran quantità di dati che sono stati raccolti per il dipartamento di Telematica del Politecnico di Torino da quando Tstat è stato creato. Tstat è uno strumento disegnato ed elaborato per lo stesso dipartamento di Telematica del Politecnico di Torino, è capace di ascoltare e catturare pacchetti che attraversano la rete ed estrarre delle informazioni utili del traffico Internet.

Il mio compito in questa tesi è stato quello di prendere questi dati raccolti, organizzarli, analizzarli ed estrarne dei risultati utili in modo da determinare parte della matrice di traffico Internet. La matrice di traffico ci dice quanti pacchetti si sono scambiati una coppia di punti di una rete. Siccome noi abbiamo soltanto un apparato di ascolto situato in un punto della rete, siamo soltanto in grado di determinare una riga ed una colonna di questa matrice, cioè il traffico che va da Torino verso gli altri POP ed il traffico che proviene dagli altri POP entrante a Torino.

Con lo studio della matrice di traffico, le aziende providers dei servizi Internet hanno una informazione che ha un grande valore per poter adattare i propri impianti e le risorse investite nella propria rete. Nel nostro caso è stato Fastweb, azienda provider di servizi Internet tramite fibra ottica, numero uno nel suo settore, che ci ha chiesto di condurre questo studio del traffico che circola atraverso la sua rete.



A partire dal livello di dettaglio che vogliamo utilizzare per investigare la rete, questa metodologia ci permette anche di cambiare la scala temporale dell'analisi o modificare l'intervallo di campioni presi in considerazione.

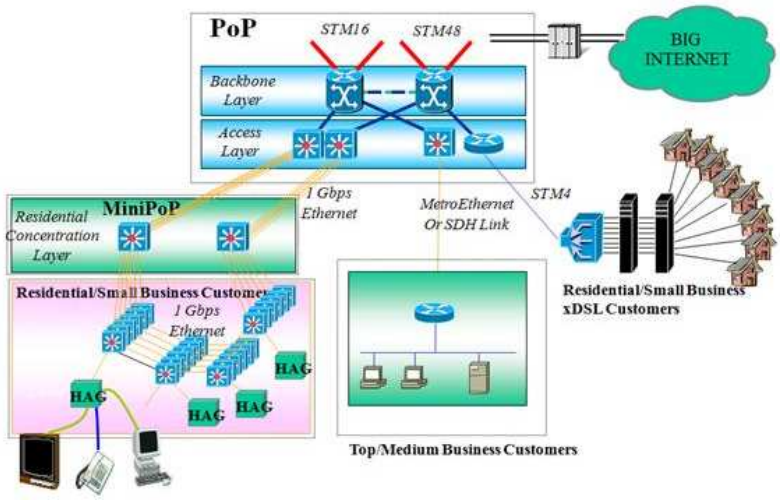
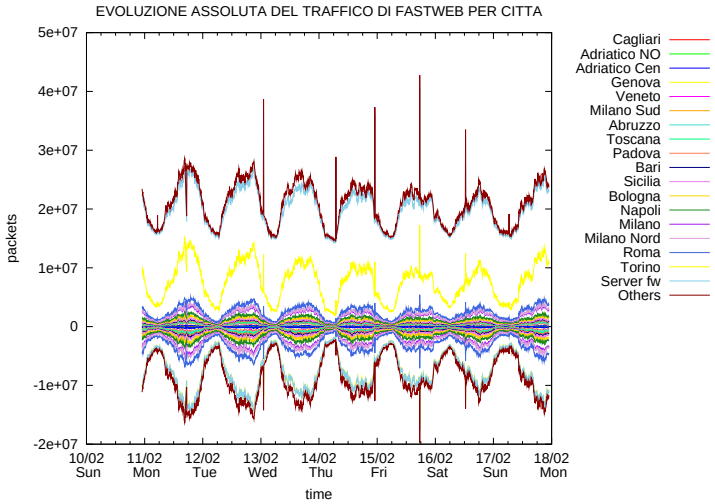


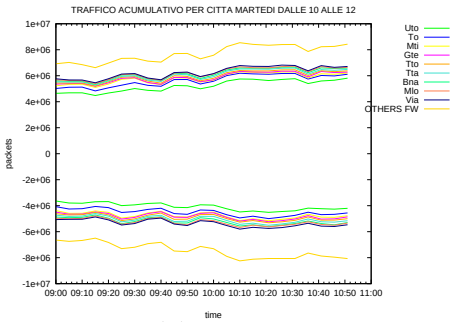
Figura 1. Architettura della rete Fastweb

Con dei file generati, e l'informazione ordinata, possiamo creare facilmente dei grafici di correlazione, per vedere quanto i traffici si assomigliano tra loro.

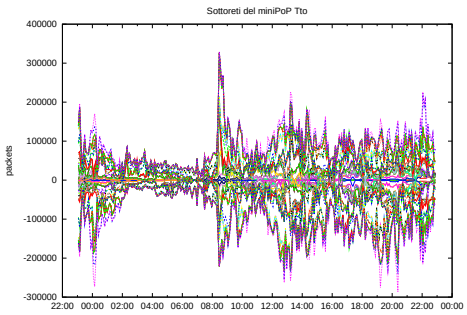
A seconda del livello di dettaglio della rete considerata, possiamo ottenere delle evoluzioni del traffico a livello di PoP, di miniPoP e anche di sottorete. Di seguito è mostrato un esempio dell'evoluzione nel tempo del traffico a livello di PoP, miniPoP e sottoreti durante diversi periodi della settimana:



(a) PoP

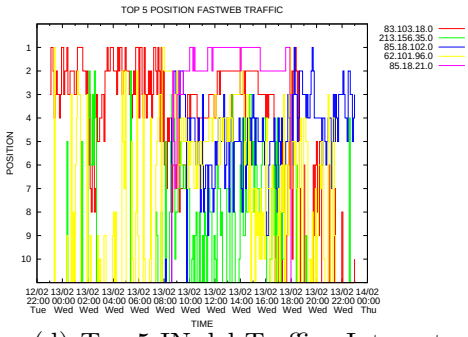


(b) MiniPoP

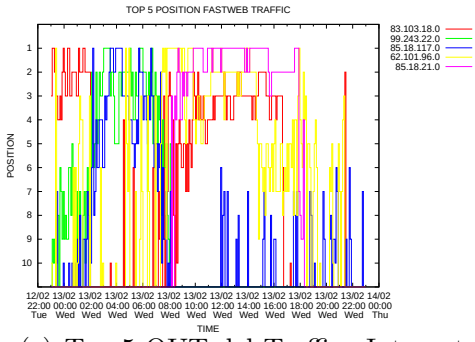


(c) Sottorete

La seconda parte dello studio ha il compito di determinare quali siti Internet sono i più visitati dagli utenti. Come nel caso precedente, possiamo scegliere il livello di dettaglio di rete che vogliamo analizzare, sicuramente però a livello di sottorete è più interessante. Ciò è evidenziato nel graffico seguente in cui possiamo vedere l'evoluzione della posizione di un gruppo d'indirizzi IP durante un giorno della settimana:



(d) Top-5 IN del Traffico Internet



(e) Top-5 OUT del Traffico Internet

I lavori da condurre in futuro possono essere orientati a creare una interfaccia grafica per la generazione automatica dei grafici allo stesso modo in cui funziona il sito http://tstat.tlc.polito.it/web.shtml

L'informazione ricavata da queste misure della rete può essere utile in un futuro studio più esteso per adattare le infrastrutture che Fastweb ha sparse per il territorio italiano, in modo da ottimizare le risorse, determinare nuovi o diversi instradamenti, etc. In definitiva, migliorare la qualità del servizio per l'utente finale in cambio di un uso più efficente della rete.

Indice

Ringraziamenti II

1 Introduzione 1
- 1.1 Motivazioni e obbiettivi . . . 1
- 1.2 Descrizione della soluzione proposta . . . 3
- 1.3 Struttura della tesi . . . 3

2 Conoscenze di base 5
- 2.1 Introduzione . . . 5
- 2.2 Fastweb . . . 5
  - 2.2.1 Architettura della rete di Fastweb . . . 5
- 2.3 Matrice di traffico . . . 7
  - 2.3.1 Metodologie di stima della matrice di traffico . . . 7
- 2.4 Sniffing Tools . . . 9
  - 2.4.1 Strumento di misura, Tstat . . . 9

3 Metodologia 11
- 3.1 Awk come linguaggio di programmazione . . . 11
- 3.2 Metodologia per l'elaborazione dell'informazione . . . 11
  - 3.2.1 Calcolo del volume di traffico . . . 11
  - 3.2.2 Elaborazione dei primi grafici dei risultati . . . 13
  - 3.2.3 Determinazione dei siti più sollecitati . . . 14
  - 3.2.4 Calcolo delle probabilità di appartenenza ai siti più visitati . . . 16

4 Analisi 18
- 4.1 Introduzione . . . 18
- 4.2 Generare la matrice di traffico di Torino. . . . 18
- 4.3 Grafici dei risultati del volume di traffico . . . 20
- 4.4 Identificare gli indirizzi IP più gettonati . . . 22
  - 4.4.1 Indentificare i siti web più visitati . . . 22
  - 4.4.2 Identificare le reti di Fastweb più visitate . . . 24

4.4.3 Come ottenere il grafico degli indirizzi IP più visitati 24

5 Risultati **26**
5.1 Introduzione 26
5.2 Distribuzione generale del traffico nella rete 27
5.3 Distribuzione del traffico per città o PoP 29
5.3.1 Correlazioni fra Torino e gli altri POP 31
5.4 Per mini-Pop 34
5.4.1 Analisi del traffico di martedì, mercoledì e sabato dalle 9 alle 11 34
5.4.2 Analisi del traffico di martedì, mercoledì e sabato dalla 1 alle 3 35
5.4.3 Analisi del traffico di giovedì dalle 7 alle 13 36
5.5 Analisi del traffico a 3 livelli, PoP, miniPoP, sottorete 38
5.5.1 Analisi del traffico verso Milano durante 24 ore 38
5.5.2 Analisi del traffico verso Roma durante 24 ore 40
5.5.3 Analisi del traffico verso Napoli durante 24 ore 41
5.6 Correlazioni giorno-notte secondo i PoP 44
5.7 Risultati dei siti più visitati 49
5.7.1 TOP5 Traffico Fastweb 49
5.7.2 TOP5 Traffico Internet 51

6 Conclusioni **54**

Bibliografia **55**

Elenco delle figure

1 Architettura della rete Fastweb IV
2.1 Architettura rete Fastweb 6
2.2 Situazione di Tstat 10
3.1 Diagramma di flusso per calcolare il volume di traffico 12
3.2 Algoritmo usato per dove.awk 12
3.3 Diagramma di flusso per disegnare il volume di traffico 14
3.4 Diagramma di flusso per calcolare i siti più gettonati 15
3.5 Algoritmo usato per topx.awk 15
3.6 Diagramma di flusso per calcolare la matrice delle probabilità che un indirizzo IP sia visitato 16
5.1 Distribuzione assoluta 27
5.2 Distribuzione relativa 28
5.3 Evoluzione assoluta traffico per città 29
5.4 Evoluzione relativa del traffico per città 30
5.5 Evoluzione assoluta del traffico per città senza il Server di Fastweb . . 31
5.6 Correlazioni tra il PoP di Torino e altri PoP della rete di Fastweb . . 32
5.7 Parametri di regressione 33
5.8 Traffico per MiniPoP dalle 9 alle 11, con la contribuzione del Server di Fastweb 34
5.9 Traffico per MiniPoP dalle 9 alle 11 34
5.10 Traffico per MiniPoP dall' 1 alle 3 35
5.11 Traffico per MiniPoP dall' 1 alle 3 36
5.12 Giovedì dalle 7 alle 13 37
5.13 Traffico secondo i PoPs 38
5.14 Milano secondo i miniPoPs 39
5.15 Traffico secondo le sottoreti 39
5.16 Traffico secondo i PoPs di Roma 40
5.17 Roma secondo i miniPoPs 41
5.18 Roma secondo le sottoreti 41
5.19 Napoli secondo i PoPs 42
5.20 Napoli secondo i miniPoPs 42

5.21 Napoli secondo le sottoreti 43
5.22 Traffico diurno totale venerdi 44
5.23 Venerdi dalle 8 alle 20 44
5.24 Venerdi dalle 8 alle 20 45
5.25 Correlazioni diurne di Torino con gli altri PoP più importanti 45
5.26 Correlazioni diurne 46
5.27 Traffico notturno totale venerdì 46
5.28 Venerdi dalle 20 alle 8 47
5.29 Venerdi dalle 20 alle 8 47
5.30 Correlazioni notturne Torino con gli altri PoP importanti 48
5.31 Correlazioni notte 48
5.32 Top 5 del Traffico Fastweb IN e OUT di Torino per tutta la settimana 49
5.33 Top 5 del Traffico Fastweb IN e OUT Torino Mercoledì 50
5.34 Top 5 del Traffico Fastweb IN e OUT Torino Domenica 50
5.35 Top 5 del Traffico Internet IN e OUT di Torino per tutta la settimana 51
5.36 Top 5 del Traffico Fastweb IN e OUT Torino Mercoledi 52
5.37 Top 5 del Traffico Fastweb IN e OUT Torino Domenica 53

Capitolo 1

Introduzione

1.1 Motivazioni e obbiettivi

L'ingegneria del traffico Internet è stata definita dal RFC[1] 2720:

"Internet traffic engineering is defined as that aspect of Internet network engineering dealing with the issue of performance evaluation and performance optimization of operational IP networks. Traffic Engineering encompasses the application of technology and scientific principles to the measurement, characterization, modelling, and control of Internet traffic."

Da questa definizione possiamo capire che l'Ingegneria del traffico è quella disciplina che si occupa che ognuno di noi abbia accesso a Internet con una buona qualità, facendo un uso responsabile ed efficiente delle risorse in modo da ottenere una comunicazione affidabile. Per ottenere questa qualità dei servizi bisogna analizzare il traffico Internet che attraverso le diverse reti per capire di che tipo di rete si tratta e applicare diverse soluzioni per rendere più efficiente la rete globale di Internet alle nostre case.

Internet usa un modello di rendimento chiamato best-effort, questo significa che non viene fornita nessuna garanzia sulla consegna dei dati o sul livello di QoS (Quality of Service), ma, tutte le comunicazioni avvengono con il massimo impegno possibile (best effort per l'appunto). La diretta conseguenza di questo approccio è rappresentata da un bitrate e un tempo di consegna variabili in base all'attuale carico della rete. Nonostante la rimozione di queste caratteristiche di controllo e

[1]Request for Comments (RFC) è un memorandum publicato dalla Internet Engineering Task Force (IETF) che descrive metodologie, comportamenti, ricerche, o innovazioni applicabili al funzionamento di Internet e ai sistemi ad esso connessi.

preallocazione delle risorse possa sembrare insensata, è in realtà utilissima, infatti, così facendo, la struttura della reti ne risulta semplificata e opera più efficientemente.

Attualmente in Internet stanno aumentando le applicazioni come video streaming, che richiedono una grande larghezza di banda o una qualità di servizio che ne garantisce il corretto funzionamento. Questo significa che si deve avere un miglior controllo della rete perché questi requisiti siano raggiunti. Anche le applicazioni peer-to-peer sono usate per scaricare una grande quantità di dati, come film o programmi, queste applicazioni tendono a occupare tutta la larghezza di banda disponibile, senza tenere in considerazione la distribuzione fatta dalla rete. La congestione è uno dei principali problemi da evitare.

Nelle reti sovradimensionate, per quello che riguarda la larghezza di banda, ci sono ancora delle parti congestionate. Alcuni siti in qualche momento possono risultare congestionati a causa di un improvviso interesse da una gran parte di clienti o anche dovuto a fluttuazioni casuali del traffico. Inoltre ci possono essere altri motivi per la variazione dell'intensità di traffico, una caduta di un link e il suo susseguente re indirizzamento provocano alte quantità di traffico in altri punti della rete. Per questi motivi l'ingegneria del traffico lavora per fare un uso corretto e adeguato delle risorse per evitare la congestione della rete.

Gli obbiettivi da tenere presente sono: la minimizzazione della perdita di pacchetti, la minimizzazione del ritardo, la massimizzazione del throughput è il raggiungimento di un accordo sulla qualità del servizio da raggiungere. A causa della politica di best-effort, perdita di pacchetti è uno dei principali obiettivi di rendimento.

Per tutto ciò ci serve un elemento matematico che ci sarà utile per affrontare questa sfida, la matrice di traffico. La matrice di traffico è un riassunto del flusso di traffico tra i punti origine e i punti destinazione della rete. E giustamente è questo quello che ci occupa in questa tesi, stimare la matrice di traffico di una rete reale ed analizzare il traffico che la attraversa.

Come si potrà vedere in successivi capitoli, data la gran difficoltà della stima della matrice di traffico con delle misure reali, noi saremo in grado di calcolare soltanto una riga e una colonna di questa matrice, che corrisponde al traffico inviato e ricevuto da Torino al resto della rete di Fastweb.

Negli ultimi anni, la stima della matrice di traffico è stato motivo di tante ricerche tra gli studiosi. Le aziende providers di servizi Internet si sono anche interessate a queste investigazioni, dato che a partire dalla sua conoscenza si possono ottimizzare

gli impianti e le risorse per pianificare la rete di tali caratteristiche, e così ridurre costi e massimizzare i profitti.

1.2 Descrizione della soluzione proposta

Il dipartimento di telematica del Politecnico di Torino, ha sviluppato un tool chiamato Tstat capace d'analizzare il traffico Internet che passa attraverso il miniPoP situato nella rete del Politecnico di Torino. Noi approfitteremo di queste tracce di dati per poterne analizzare il traffico che entra ed esce da questo miniPop però a livello IP, guardando i pacchetti che entrano ed escono tanto a livello di Pop, miniPop ed anche a livello di sottorete. Con tutta questa informazione proveremo a generare dei grafici utili per capire la distribuzione del traffico della rete e saremo anche in grado di calcolare una riga ed una colonna della matrice di traffico. In una seconda parte dello studio si farà un'analisi dei siti web più visitati e la probabilità di essere visitati, tutto questo in funzione del giorno della settimana e delle ore diurne o notturne.

Siccome lavoreremo con una gran quantità di dati, la maniera più efficiente e facile per lavorare sarà svilupando dei piccoli programmi che ci aiuteranno a gestire e organizzare i dati per poi estrarne delle conclusioni. La programmazione si può portare a termine con qualsiasi linguaggio, però sarà più utile usare AWK che serve ad applicare a una gran quantità di dati o di file lo stesso programma. Gnuplot sarà il programma che si occuperà di mostrarci i grafici risultanti. E in qualche occasione ci potremmo servire di Matlab, un altro tool utilissimo per la computazione matematica e statistica.

Questo lavoro intende stabilire una metodologia standard tramite la quale in un futuro si sia in grado di analizzare qualsiasi insieme di risultati catturati per uno sniffer come Tstat e ottenerne dei grafici utili per ulteriori ridimensionamenti di risorse o una pianificazione più efficace di ogni PoP della rete.

1.3 Struttura della tesi

Questo documento è organizato il tre parti differenziate, una prima parte dove si parlerà delle conoscenze di base di cui abbiamo bisogno per capire l'ambiente in cui stiamo lavorando. In questo caso faremo una descrizione della rete di Fastweb, rete di cui stiamo analizando il traffico. Un altro elemento da prendere in considerazione è Tstat, il tool creato dal Politecnico di Torino che ci fornisce i dati per la sua

sucessiva analizazzione. Il terzo punto delle conoscenze di base intende essere un piccolo sommario degli studi che sono stati effettuati precedentemente a questa tesi per quanto riguarda la stima della matrice di traffico.

Dopo aver descritto l'obbiettivo della tesi, si porterà a termine l'analisi e la metodologia usata per fare questa analisi: i linguaggi di programmazione, i tools impiegati in tutto il processo e come sono stati usati.

In fine nella terza ed ultima parte vengono commentati i risultati grafici ottenuti dopo l'analisi del traffico e la stima della matrice di traffico. Le conclusioni saranno presenti anche nell'ultima parte del documento.

Capitolo 2

Conoscenze di base

2.1 Introduzione

In questa sezione si svilupperano tre punti fondamentali per comprendere cosa si intende fare, il come e il perchè. Siccome il nostro scopo è trovare una riga e una colonna della matrice di traffico internet, prima dovremo sapere come si costruisce questa matrice e cosa significa. Il secondo punto è capire su che tipo di rete stiamo calcolando questa matrice, la rete di accesso a Internet di Fastweb. Per finire, nel terzo punto parleremo del programma che si è usato per catturare il traffico di dati.

2.2 Fastweb

Fastweb è un'azienda italiana di comunicazioni specializzata nella telefonia terrestre, nelle connessioni a banda larga e nella televisione via cavo è la più importante azienda italiana delle comunicazioni in fibra ottica. Fastweb è nata nell'ottobre del 1999 come una idea rivoluzionaria di fornire accesso a Internet agli utenti, tanto al singolo cliente quanto alle grande aziende offrendo servizi di telecomunicazioni su reti IP. Nell'ottobre del 2000, il servizio fu aperto ai clienti e anche alle aziende, offrendo accesso a Internet, telefonia VoIP e servizi di video on demand. Da quel momento Fastweb è diventato la maggiore azienda di telecomunicazioni a banda larga in Italia.

2.2.1 Architettura della rete di Fastweb

Grazie alla sua architettura IP e alla capacità di usare indistintamente le tecnologie di FTTH come xDSL per accedere a Internet, Fastweb ha ottimizzato l'integrazione di servizi di dati, VoIP o IPTV su una sola connessione a banda larga.

In questa sezione introdurremo brevemente l'architettura di Fastweb.

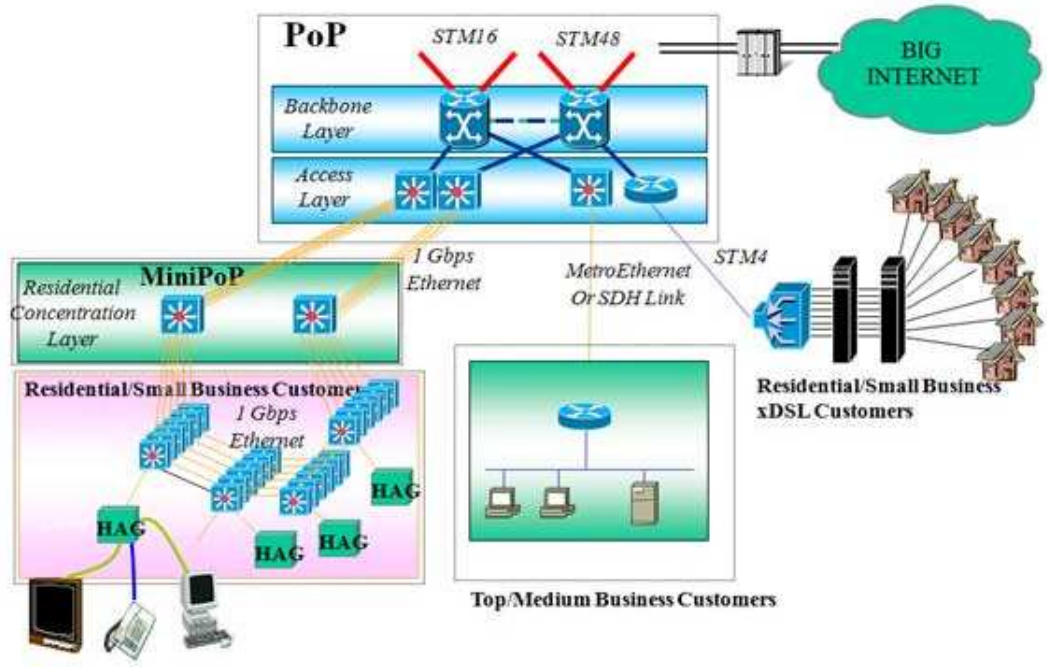


Figura 2.1. Architettura rete Fastweb

Come si rappresenta nella Figura 2.1, nell'ultima fase è adottata una architettura Ethernet basata su una MAN (rete di area metropolitana). Nelle residenze e piccole aziende, i clienti sono collegati tramite un HAG (Home Acces Gateway) il quale è in grado di offrire porte Ethernet per collegare PC, i Video Box ed anche i tradizionali telefoni. Un HAG è essenzialmente un Switch Ethernet, combinato con un gateway H.323 per convertire il segnale analogico d'ingresso nel POT in una segnale di trasporto VoIP. Per collegare l'HAG ad un switch viene usata una porta 10-Base-F se si ha un accesso FTTH (Fiber-To-The-Home); viene, invece, usato un modem se si ha un accesso xDSL.

Le città più grandi in Italia sono direttamente interconnesse con più di 12.400Km di fibra ottica. In ogni città, sono presenti uno o più PoPs, mentre tanti Mini PoPs sono usati per raccogliere il traffico di fino a 10.000 utenti.

Nel caso trattato, noi abbiamo messo uno sniffer di pacchetti IP in uno di questo Mini PoPs di Torino.

2.3 Matrice di traffico

La matrice di traffico di una rete di telecomunicazioni, come potrebbe essere una rete d'accesso a Internet, è una espressione matematica utilizzata per avere insieme l'informazione del flusso di dati che si scambiano tra due punti della rete. Questi due punti sono denominati punto origine e punto destinazione. Nel caso analizzato viene preso in considerazione uno dei Mini Pop di Torino verso gli altri PoP della intera rete Fastweb. La matrice di traffico ha la seguente forma:

$$\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix}$$

Dove l'elemento $a_{m,n}$ rappresenta il traffico originato dal punto m e con punto destinazione n. Questi punti $a_{m,n}$ possono essere a qualsiasi livello, tanto Pop quato di miniPop o sottorete. Va comunque tenuta in considerazione l'aumento di difficoltà computazionali e di analisi che si incontrano a livelli inferiori.

2.3.1 Metodologie di stima della matrice di traffico

Ci sono tanti metodi per la stima proposti in letteratura, la maggior parte dei quali possono essere classificati in tre grandi gruppi a seconda di come sono le supposizioni di partenza. Queste sono:

- Modello gravitazionale [3]

  Il modello gravitazionale suppone che il traffico tra due punti è proporzionale al traffico totale originato tra il nodo sorgente e il nodo destinazione. Questa supposizione implica che non ci sono nodi nella rete che comunicano tra loro più del traffico totale supposto. Se ciò è sempre verificato, si può ottenere una stima iniziale precisa usando soltanto il traffico totale che esce ed entra in ogni nodo. Questa potrebbe essere utilizzata assieme al conto dei link, per ottenere la stima finale.

- Rapporto media-varianza [4, 5]

  Per quanto riguarda il rapporto media-varianza, si suppone che la varianza del flusso di traffico dipende della media di volume del flusso di traffico mediante una funzione di legge potenziale. Questo implica che quanto più grande

è il flusso di traffico, ci sono anche delle maggiori variazioni del traffico. Si può quindi usare la varianza tra le coppie origine e destinazione per stimare la media tra le coppie origine e destinazione di traffico. Se fosse avessimo delle misure, sarebbe possibile ottenere campioni di covarianze del carico dei link. Se si parte delle covarianze, è possibile trovare le varianze che servono per trovare le medie del traffico delle coppie origine-destinazione basandosi nel rapporto media-varianza. Tanto il primo momento (il carico dei links) quanto il secondo momento (la covarianza del carico dei links) delle misure sono usati. Metodi di questo tipo sono chiamati metodi di secondo momento.

- Misure dirette [6]

  C'è un terzo gruppo di metodi di stima anche chiamati metodi di terza generazione. Sono fondamentalmente diversi degli altri gruppi di metodi in quanto non si presuppone che le misure non siano disponibili. Invece, si basa nella presupposizione che si possono ottenere le misure. Il problema a risolvere diventa ottenere un compromesso di minimizzazione dell'overhead ed il tempo di misura, però allo stesso tempo mantenere l'errore di stima ancora piccolo.

  La misurazione e le caratteristiche del traffico sono strettamente legate alla problematica della stima della matrice di traffico. Per essere ingrado di effettuare mansioni dell'ingegneria del traffico, come il bilanciamento di carico, abbiamo bisogno di avere prima delle misure del traffico della rete e così saperne il suo stato. La misurazione del traffico è una parte importante nella ingegneria del traffico in un altro aspetto ancora, le misure possono essere usate dentro e fuori l'analisi per caratterizzare il traffico internet in generale. In particolare, si possono usare per testare delle ipotesi usate per stimare la matrice di traffico ed analizzare la validità dei modeli gravitazionale e quello del rapporto media-varianza.

2.4 Sniffing Tools

Si definisce *sniffing* l'attività di intercettazione passiva dei dati che transitano in una rete telematica. Tale attività può essere svolta sia per scopi legittimi (ad esempio l'individuazione di problemi di comunicazione o di tentativi di intrusione) sia per scopi illeciti (intercettazione fraudolenta di password o altre informazioni sensibili).

I prodotti software utilizzati per eseguire queste attività vengono detti sniffer ed oltre ad intercettare e memorizzare il traffico offrono funzionalità di analisi del traffico stesso.

Gli sniffer intercettano i singoli pacchetti, decodificando le varie intestazioni di livello datalink, rete, trasporto, applicativo. Inoltre possono offrire strumenti che analizzano ad esempio tutti i pacchetti di una connessione TCP per valutare il comportamento del protocollo o per ricostruire lo scambio di dati tra le applicazioni.

2.4.1 Strumento di misura, Tstat

Tstat è un tool sviluppato dal Gruppo Reti di Telecomunicazioni del Politecnico di Torino, Tstat è un software open source che è in continuo sviluppo e si presenta come uno strumento flessibile, in grado di analizzare il traffico TCP/IP e di ricavare indici prestazionali non solo a livello rete (IP), ma anche a livello trasporto (TCP). Questo secondo aspetto permette di ricavare indici prestazionali direttamente sensibili dall'utente della rete, come ad esempio il throughput o il tempo di download di un file o di una pagina web.

Tstat sfrutta librerie standard di software e offre agli amministratori di reti e ricercatori importanti informazioni su classici e nuovi indici di attuazione e dati statistici sul traffico Internet.

Nato come una evoluzione di TCPtrace, Tstat analizza tracce di pacchetti real-time catturate usando un PC, o anche tracce precedentemente registrate in diversi formati dump (ad esempio quelle supportate per la libreria libpcap [7] , e i sistemi DAG [8] usati per monitorare links di velocit pari ai Gigabit.) È scritto in linguaggio C standard, funziona su sistemi Linux, e dovrebbe funzionare anche su altri sistemi Unix.

Oltre a generare statistiche IP, derivate dall'analisi dell'intestazione del livello IP, Tstat ricostruisce ogni stato di connessione TCP analizzando l'intestazione TCP nei flussi forward e backward di pacchetti.

Tstat costruisce istogrammi di indici misurati, scaricando la distribuzione ottenuta periodicamente (in intervalli di 5min) e non ad ogni singolo dato misurato. I dati generati tramite l'analisi statistico può essere visualizzato on-line. È in grado di calcolare più di 80 tipi diversi di istogrammi, effettuare statistiche a livello TCP e IP. Queste vengono fatte sia su misure facilmente ottenibile direttamente dagli intestazioni dei pacchetti (percentuale di pacchetti TCP o UDP, distribuzione della lunghezza dei pacchetti, distribuzione delle porte TCP...) sia su misure più complesse per quanto riguarda a TCP (dimensione media della finestra di congestione, stima del RTT, data out-of-sequence, data duplicata...).

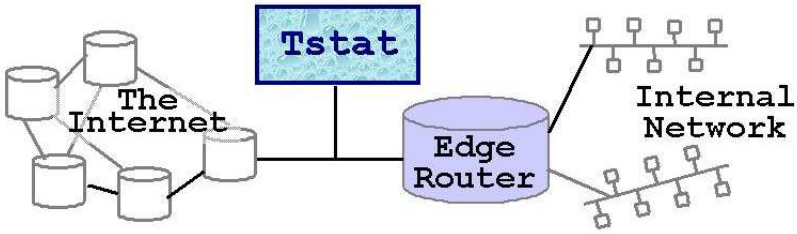


Figura 2.2. Situazione di Tstat

Nella Figura 2.2 possiamo vedere dove deve essere inserito Tstat dentro la rete per ottenere dei pacchetti.

In particolare noi siamo interessati ad un insieme di files, generati ogni 5 minuti, che contengono indirizzi IP e pacchetti inviati e ricevuti da questo indirizzo. Questa sarà la informazione basica e necessaria per sviluppare il nostro lavoro.

Capitolo 3

Metodologia

3.1 Awk come linguaggio di programmazione

Tutti i dati estratti da Tstat vengono analizzati usando dei programmi creati ad-hoc per trattare questa informazione. Data la grossa mole di dati in nostro possesso si è utilizzato AWK come linguaggio di programmazione. AWK un linguaggio adatto ad eseguire un comando su tutte le righe di un file. Nel nostro caso si tratta del tool perfetto per il nostro lavoro di analisi.

3.2 Metodologia per l'elaborazione dell'informazione

3.2.1 Calcolo del volume di traffico

Dove.awk è il programma che si occupa di prendere tutti i file di *addresses* che contengono l'informazione dei pacchetti inviati e ricevuti da un indirizzo IP, per generare il file *results.txt*. Tutti i percorsi per arrivare fino ai file *addresses* sono contenuti dentro il file *directories.txt*.

Per eseguire il programma *dove.awk* ci serve il file di configurazione per le maschere di rete su cui vogliamo avere classificati i risultati, *netmask.txt*. Per avere ulteriori informazione relative ai formati dei file, si rimanda al capitolo 4. Nella figura 3.1 possiamo vedere il diagramma di flusso di lavoro del programma *dove.awk*.

La prima cosa di cui si occupa il programma *dove.awk* è di caricare la configurazione delle maschere di rete. Per ogni file di *addresses*, *dove.awk* analizza ogni riga del file e prende la colonna dove si trovano gli indirizzi IP. Per ogni indirizzo IP si fa una maschera bit a bit con tutti gli indirizzi di rete del file di *netmask.txt*.

Figura 3.1. Diagramma di flusso per calcolare il volume di traffico

Quando si trova la rete che coincide alla maschera, si aggiorna la somma di pacchetti inviati e ricevuti dalla rete corrispondente. Possimao trovare i pacchetti inviati nella seconda colonna e i pacchetti ricevuti nella terza colonna del file di *addresses*. Dopo l'elaborazione di questi dati i risultati vengono inseriti in un file di risultati, *results.txt*, specificando per ogni istante di tempo quanti pacchetti sono stati inviati e ricevuti da ogni rete. L'istante di tempo è contenuto nel nome del percorso dove si trova ogni file di *addresses*, perché questo percorso ha nascosto in se stesso la data e l'ora in cui ogni file di *addresses* è stato generato da Tstat.

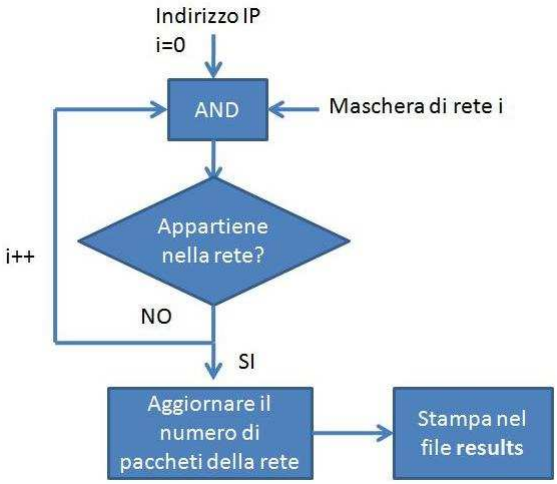


Figura 3.2. Algoritmo usato per dove.awk

Nel file di *results.txt* si trovano organizzati per data e ora, i pacchetti che sono stati ricevuti ed inviati per ogni rete, questo per tutti i pacchetti che sono stati catturati da Tstat. Il programma *dove.awk* può essere eseguito per ogni livello di

dettaglio del traffico da analizzare (PoP, mini-PoP o sottorete). In ognuno dei casi descritti si ha bisogno di specificare il corrispondente file di configurazione, *netmask.txt*.

3.2.2 Elaborazione dei primi grafici dei risultati

Dopo aver eseguito il programma *dove.awk* ed ottenuto il file di risultati, *results.txt*, si deve fare una piccola elaborazione dei dati prima di generare i grafici. Nel file di risultati vengono enumerati i diversi PoP o miniPoP col loro corrispondente traffico d'ingresso e di uscita; inoltre viene anche enumerato il traffico corrispondente a Multicast e altri siti che non appartengono nella rete di Fastweb. Perciò, prima di creare i grafici relativi al traffico Fastweb dobbiamo liberarci delle colonne relative al traffico di Multicast ed Others con un semplice comando Unix.

Dobbiamo individuare le colonne che occupano i traffici da cancellare e poi lanciare il comando *cut -dx,y,z* e così avremo soltanto quelli che vogliamo rappresentare graficamente (x,y,z sono le colonne che devono rimanere). Se invece non togliamo il traffico di Multicast, nei nostri graffici verrà ponderato ogni traffico con la somma totale di traffici, compreso il contributo di Multicast, ed è proprio quello che vogliamo evitare.

Rielaborato il file *results.txt*, dobbiamo accertarci di avere anche il file di configurazione delle maschere di rete che vogliamo che compaiono nel grafico, *netmask_to_plot.txt* e anche l'intervallo di campionamento. Quest'ultima informazione la possiamo trovare nel file *data_to_plot.txt*.

Il programma *read.awk* legge dal file *netmask_to_plot.txt* le reti di cui vogliamo fare i nostri grafici e le tiene in memoria. Dopo di che legge il file *data_to_plot.txt* che contiene l'istante iniziale e finale del grafico. Di seguito prende ogni colonna di dati del file di risultati e controlla se il nome della colonna corrisponde a quelli che ha salvato in memoria. Se la rete considerata è contenuta nel file di reti da rappresentare ed è compresa nel periodo indicato per *data_to_plot.txt*, allora verrà rappresentata nel grafico. Il quinto parametro del file di *data_to_plot.txt* contiene l'intervallo dei campioni, se questo parametroè superiore a 5 minuti, o 300 secondi, si sommeranno i pacchetti contenuti nell'intervallo e verranno disegnati come se fossero il contributo di un solo campione.

L'esecuzione del programma *read.awk* produce due grafici relativi al volume del traffico (l'assoluto ed il relativo). Con la dovuta elaborazione dei file di risultati,

dei file *netmask_to_plot.txt* e *data_to_plot.txt* abbiamo ottenuto i grafici che contengono soltanto le reti specificate in *netmask_to_plot.txt* nell'intervallo indicato in *data_to_plot.txt*. Questi risultati sono analizzati nelle sezioni 5.3 e 5.4 del capitolo 5.

Anche in questo caso, mostriamo il diagramma di flusso, con i file di input necessari ad ottenere i grafici nella Figura 3.3 ed i suoi corrispondenti file di output.

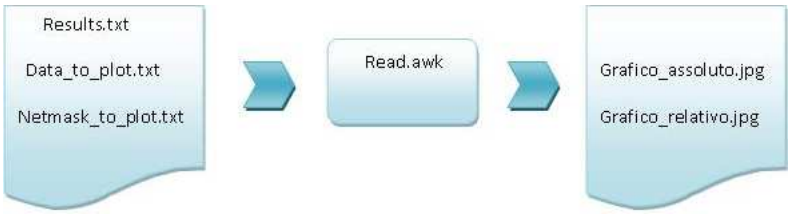


Figura 3.3. Diagramma di flusso per disegnare il volume di traffico

3.2.3 Determinazione dei siti più sollecitati

Come è illustrato nei paragrafi precedenti, *directories.txt* contiene il percorso dei file di *addresses* che contengono le informazioni da analizzare, mentre in questo caso *netmask.txt* contiene il file delle maschere di rete sulle quali vogliamo calcolare i siti più visitati.

Ci sono due versioni del programma *topx.awk*, una che calcola soltanto il ranking degli indirizzi IP compresi nel file delle maschere, ed un'altra versione *topx_fw.awk*, che calcola il ranking degli indirizzi IP ad esclusione di quelli compresi nel file delle maschere. *Topx.awk* è stato creato pensando di fare una classifica soltanto di un gruppo di reti. Invece *topx_fw.awk* è stato creato pensando di fare la classifica del traffico che resta dopo aver tolto il traffico che a noi noto. In base al tipo di grafico che vogliamo analizzare useremo una versione o l'altra.

Entrambe le versioni hanno come risultato due file. Uno contiene il ranking degli indirizzi d'ingresso e l'altro di quelli d'uscita. Il file viene organizzato nel seguente modo: per ogni istante di tempo indica quali indirizzi IP sono stati più gettonati al primo posto, secondo, etc... ed il numero dei pacchetti che sono stati registrati per quegli indirizzi IP. All'inizio del programma c'è un parametro da configurare per dire al programma quanti posti vogliamo nel nostro ranking, la 'x' del TOPx.

Figura 3.4. Diagramma di flusso per calcolare i siti più gettonati

Il programma *topx.awk* prende ogni file di *addresses* e per ogni rete non contenuta nel file di *netmask.txt* somma i pacchetti che entrano e che escono dalla rete. Dopo di che, li mette in ordine decrescente e in fine prende gli 'x' primi posti di questo file e li stampa per colonne nei file TOPx_IN e TOPx_OUT rispettivamente.

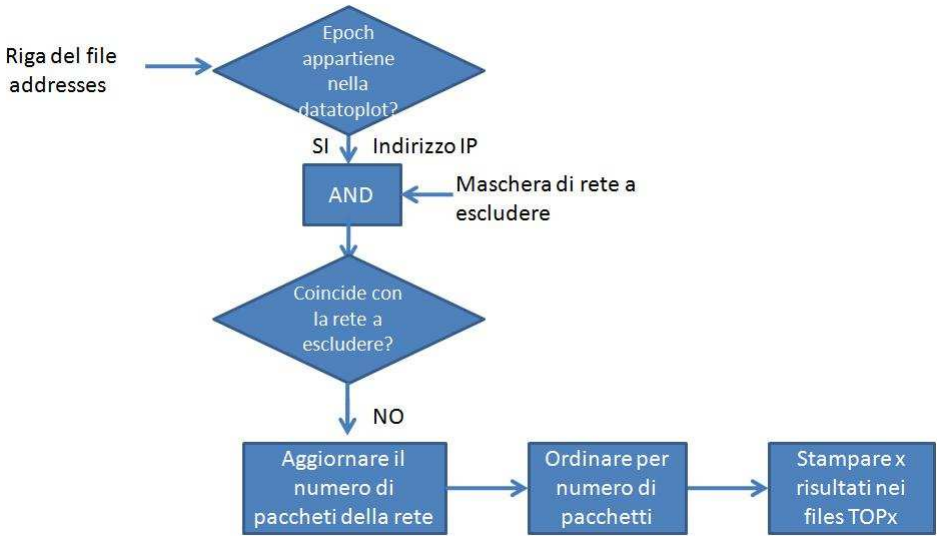


Figura 3.5. Algoritmo usato per topx.awk

TOPx_IN e TOPx_OUT contengono per ogni istante di tempo l'indirizzo IP che sta al primo posto del ranking e il numero di pacchetti che si ha registrato, lo stesso accade nel secondo posto del ranking, e così via fino al massimo indicato per 'x'.

3.2.4 Calcolo delle probabilità di appartenenza ai siti più visitati

Dopo aver generato i file TOP_IN e TOP_OUT, ottenuti eseguendo il programma *topx.awk* del paragrafo precedente, disponiamo dei file necessarii per concludere lo studio che stiamo faccendo. Di seguito il diagramma di flusso utilizzato per calcolare la matrice delle probabilità che un indirizzo IP sia visitato:

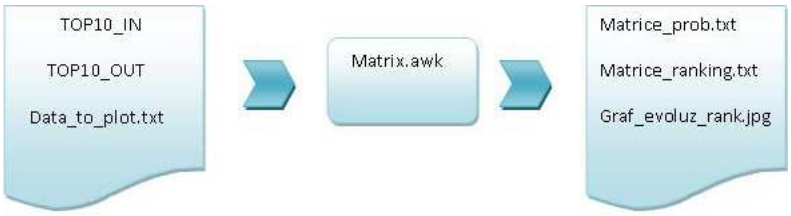


Figura 3.6. Diagramma di flusso per calcolare la matrice delle probabilità che un indirizzo IP sia visitato

Effettivamente abbiamo già calcolato i due file di input necessari, mentre il terzo file di cui abbiamo bisogno è stato già usato, *data_to_plot.txt*. Questo file contiene l'istante iniziale e l'istante finale che vogliamo rappresentare. Dobbiamo avere presente che adesso, a differenza del caso precedente, non possiamo scegliere l'intervallo dei campioni delle misure.

L'ultimo programma che useremo, *matrix.awk*, si occupa di elaborare i file TOPx_IN e TOPx_OUT. Per tutti gli IP che sono stati rilevati ci dice che posizione occupano in un ranking che ha come massimo 'x'. Questo parametro è stato configurato precedentemente per calcolare i TOPx_IN e TOPx_OUT. Se l'indirizzo IP che stiamo considerando occupa una posizione inferiore a 'x', questo indirizzo prenderà come posizione del ranking il valore di 'x+1'. Questo ci aiuterà nel momento in cui disegneremo dei grafici. Il file che contiene tutti questi dati è *matrice_ranking.txt*

Per ogni riga del file TOPx_IN o TOPx_OUT, *matrix.awk* crea una matrice i cui indici sono l'indirizzo IP e l'istante di tempo. In questa posizione della matrice si mette la posizione del ranking che l'indirizzo occupa in questo istante di tempo. Praticamente si genera una matrice di dimensione x per tanti istanti di tempo calcolati. *Matrice_ranking.txt* contiene per ogni istante di tempo, la posizione che occupano tutte le reti presenti nel file TOPx_IN o TOPx_OUT, in base al file usato come input.

Il secondo file che crea il programma *matrix.awk*, è chiamato *matrice_prob.txt*. Contiene la probabilità che ogni indirizzo IP sia nella prima posizione del ranking, nella seconda, e così via fino al parametro 'x'. L'ultima colonna mostra la probabilità di appartenere al TOP-x, essendo 'x' anche il parametro precedentemente configurato in *topx.awk*.

In un secondo momento, *matrice.awk* percorre tutta la matrice che contiene le posizioni del ranking di ogni indirizzo. Se il suo contenuto non è vuoto, si crea la matrice di probabilità i cui indici sono l'indirizzo IP e la posizione occupata nel ranking; il contenuto invece è la somma di tutte le volte che un dato indirizzo IP occupa una certa posizione nel ranking. Quindi si sommano tutte le diverse probabilità e si dividono per il numero di istanti di tempo calcolati. Infine si mettono nel file *matrice_prob.txt* tutti gli indirizzi IP con la probabilità per ognuno di essere nella prima, seconda, etc... posizione; l'ultima colonna sarà la somma di tutte le precedenti, poichè è la probabilità di essere nel TOPx.

Con *matrice_prob.txt* e *data_to_plot_topx.txt* saremo in grado di disegnare l'ultimo gruppo di grafici. Nel programma *matrix.awk* possiamo configurare il parametro 'y' che sarà il numero d'indirizzi IP che vogliamo che compaiono nel grafico.

Capitolo 4

Analisi

4.1 Introduzione

In questo capitolo si parlerà in profondità dei file di cui si ha bisogno, del formato che devono avere ed la forma giusta di eseguire ogni comando per tale di ottenere i grafici desiderati.

4.2 Generare la matrice di traffico di Torino.

In anzitutto abbiamo bisogno di avere organizzato in un file tutti i percorsi di dove si trovono i files di addresses. Questo file lo chiameremo *directories.txt* e viene organizzato della seguente maniera:

```
./14_35_10_Feb_2008.out/000/addresses
./14_35_10_Feb_2008.out/LAST/addresses
./14_41_10_Feb_2008.out/000/addresses
./14_41_10_Feb_2008.out/001/addresses
./14_41_10_Feb_2008.out/002/addresses
./14_41_10_Feb_2008.out/003/addresses
./14_41_10_Feb_2008.out/004/addresses
...
```

Il programa *dove.awk* somma e calcola tutti i pacchetti che escono ed entrano a Torino provenienti dai diversi nodi della rete. Prima di eseguirlo dobbiamo essere sicuri di essere nel direttorio giusto e anche che ci sia il file di configurazione delle

maschere di rete. Quando è tutto pronto possiamo eseguire il comando seguente:

$ ./dove.awk ‘cat directories.txt‘

Se invece vogliamo specificare il file di netmask diverso a quello che c'è già nel direttorio, lo possiamo specificare col comando:

$ ./dove.awk -v netmask_file ‘cat directories.txt‘

Il file che contiene le maschere di rete deve avere il seguente formato:

#IP address	netmask	city
1.0.0.0	8	Milano
2.0.0.0	8	Milano Hinterland Nord
5.0.0.0	8	Genova
10.0.0.0	8	Server Fastweb
11.0.0.0	8	Veneto
14.0.0.0	8	Milano Hinterland Sud
23.0.0.0	8	Roma
27.0.0.0	8	Toscana
29.0.0.0	8	Padova
31.0.0.0	8	Bari
36.0.0.0	8	Sicilia
37.0.0.0	8	Bologna
38.0.0.0	8	Torino
39.0.0.0	8	Napoli
224.0.0.0	4	Multicast
0.0.0.0	0	Other traffic

Entrambi comandi generano un file *results.txt* dove possiamo incontrare organizzato per giorno ed ora il traffico che entra ed esce da ogni PoP, dipendendo di come si abbia deffinito nel file di *netmask.txt*

Il file di *results.txt* risulterà nel formato che ci mostra la figura a continuazione:

```
#1.0.0.0        8       Milano
#5.0.0.0        8       Genova
#38.0.0.0       8       Torino
...

#Epoch      Ora     Data          Milano-IN   Milano-OUT   Genova-IN   Genova-OUT
1202651160  14:46   10/02/2008    992341      1021272      242418      336078
1202651460  14:51   10/02/2008    985505      1035627      215493      311303
...
```

Possiamo vedere, che al inizio del file c'è una legenda delle maschere e a continuazione, vengono organizati per ogni intervallo di tempo, quanti pacchetti sono stati scambiati tra Torino e gli altri PoP della rete.

4.3 Grafici dei risultati del volume di traffico

Con i seguenti documenti: *results.txt*, *netmask_to_plot.txt*, *data_to_plot.txt* e *style.sty* possiamo eseguire il progrma *read.awk* così:

$./read.awk ./results.txt

- *read.awk* è il programma che si occupa di leggere il file generato precedentemente *results.txt*, e con questa informazione è capace di creare i grafici.
- Nel file *netmask_to_plot.txt* si indicano soltanto quelle città che vogliamo che appaiono nei grafici. Le altre saranno ignorate.

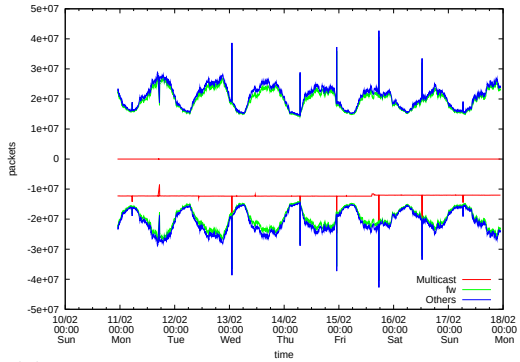
#IP address	Netmask	City
1.0.0.0	8	Milano
5.0.0.0	8	Genova
23.0.0.0	8	Roma
39.0.0.0	8	Napoli
28.0.0.0	8	Torino

- Il file *data_to_plot.txt* contiene i tempi d'inizio, tempo finale e l'intervallo per disegnare i grafici, con questo esatto formato:

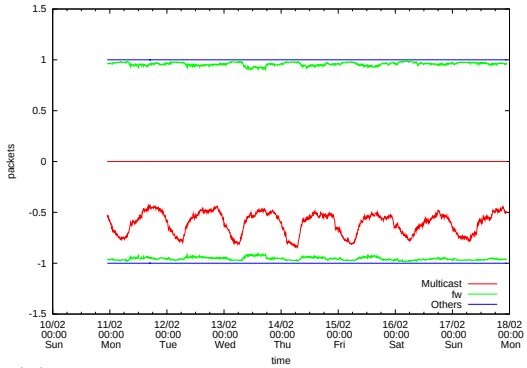
#time_start	date_start	time_end	date_end	interval
#HH:MM	DD/MM/YYYY	HH:MM	DD/MM/YYYY	(seconds)
22:20	12/02/2008	22:50	13/02/2008	3600

- E in fine c'è *style.sty*. È un file di configurazione che deve essere nello stesso directorio che gli altri documenti per avere un formato e i colori specifici nei grafici.

I due file generati si mostrano a continuazione. Nel grafico della Figura 4.1(a) possiamo vedere l'evoluzione assoluta accumulativa dei pacchetti inviati e ricevuti per le diverse città durante un intervallo di tempo di una settimana, mentre che invece, nel grafico 4.1(b) possiamo vedere l'evoluzione relativa accumulativa nello stesso periodo di tempo.



(a) Evoluzione assoluta dei principali PoP



(b) Evoluzione relativa dei principali PoP

4.4 Identificare gli indirizzi IP più gettonati

4.4.1 Indentificare i siti web più visitati

I TOP-10 è stato creato per avere evidenza dei siti più gettonati, la classifica viene realizzata dal programma *topx.awk*. La maniera per eseguire il programma è la seguente:

$ ./topx.awk ‘cat directories.txt‘

Nello stesso folder ci deve essere anche il file di *netmask_exclude.txt*, col seguente formato:

#IP address	Netmask	City
10.0.0.0	8	Server_fw
224.0.0.0	4	Multicast

Nel file di *netmask_exclude.txt* devono esserci tutti quelli gruppi d'indirizzi IP che non vogliamo considerare per calcolare il ranking d'indirizzi più gettonati, sia perché sono indirizzi IP interni di Fastweb o per qualsiasi altro motivo. Dipendendo del numero di siti che vogliamo nel nostro ranking, si deve modificare la variabile x nel file *topx.awk*, cioè:

```
BEGIN{
if (conf_file == "")
        conf_file = "netmask_exclude.txt";
load_conf_file(conf_file);

%Introduce the x from the TOPX
x=3;
}
```

In questo esempio di codice si generarà un ranking TOP3.
Come risultato di eseguire questo programma, vendrano fuori 2 files chiamati *TOP3_IN* e *TOP3_OUT*. Ognuno di loro contiene la lista degli IP's in ogni instante di tempo di cui si riceve più pacchetti (TOP_IN) o quelle più richieste (TOP_OUT). In realtà entrambe sono dipendenti l'una dell'altra.
Vediamo un essempio:

TOPX_IN:								
#Epoch	Ora	Data	1-IP	PACK	2-IP	PACK	3-IP	PACK
1202650860	14:41	10/02/08	85.14.218.0	96632	87.255.33.0	40221	85.17.132.0	39477
1202651160	14:46	10/02/08	85.14.218.0	95108	64.72.115.0	47542	208.49.82.0	42123
1202651460	14:51	10/02/08	85.14.218.0	95297	87.106.15.0	55001	42.243.97.0	47652
1202651760	14:56	10/02/08	87.106.15.0	52343	85.14.218.0	48883	208.49.82.0	40108
...								

Una volta che abbiamo ottenuto questi files, gli useremo come input del programma *matrice.awk* per calcolare la matrice di probabilità di essere in prima posizione, seconda, etc...
Il file di output *matrice_prob.txt* contiene le probabilità calcolate con il programma *matrice.awk*. In oltre viene generato il file *matrice_ranking.txt*, che contiene l'evoluzione temporale della posizione che occupano i siti più visitati. Analogamente per TOP_OUT.

Ecco come eseguire il programma:

$ ./matrix.awk TOP_IN

Dobbiamo assicurarci di avere dentro lo stesso direttorio i files che si usano per eseguire *matrix.awk*, cioè *data_to_plot.txt*, che contiene il tempo d'inizio e di fine della ricerca del ranking.

Le probabilità di ogni indirizzo IP di essere nel primo posto, secondo posto, etc..., e in fine di essere nel TOPx ad un dato instante si troveranno del file *matrice_prob.txt*, che ha la seguente struttura:

#IP	prob.1stplace	prob.2ndplace	prob.3rdplace	prob.top3
# 87.255.33.0	0.000000	0.111111	0.000000	0.111111
# 208.49.82.0	0.000000	0.000000	0.222222	0.222222
# 85.14.218.0	0.888889	0.111111	0.000000	1.000000
# 42.243.97.0	0.000000	0.444444	0.222222	0.666667
...				

Ad ogni instante possiamo vedere in che posizione si trovano gli IPs più visitati nel file *matrice_TOPX_IN.txt*

#EPOC	IP	Rank	IP	Rank	IP	Rank	IP	Rank
1202650860	87.255.33.0	2	208.49.82.0	4	85.14.218.0	1	42.243.97.0	4
1202651160	87.255.33.0	4	208.49.82.0	3	85.14.218.0	1	42.243.97.0	4
1202651460	87.255.33.0	4	208.49.82.0	4	85.14.218.0	1	42.243.97.0	3
1202651760	87.255.33.0	4	208.49.82.0	3	85.14.218.0	2	42.243.97.0	4
1202652060	87.255.33.0	4	208.49.82.0	4	85.14.218.0	1	42.243.97.0	2
1202652360	87.255.33.0	4	208.49.82.0	4	85.14.218.0	1	42.243.97.0	3
1202652660	87.255.33.0	4	208.49.82.0	4	85.14.218.0	1	42.243.97.0	2

Per semplificare la lettura del grafici, agli indirizzi IP che non sono nella posizione Topx ad un dato instante di tempo, viene assegnata la posizione x+1 (4 nel caso in esame). Vedere nella figura sopra di queste linee.

Analogamente possiamo ripetere tutto il processo per il file TOP_OUT:

$ ./matrix.awk ./TOPX_OUT

4.4.2 Identificare le reti di Fastweb più visitate

Per eseguire il programma si lancia la seguente istruzione:

$ ./topx_fastweb.awk netmask_include.txt

Il programma è molto simile a *topx.awk*, però con una sola differenza, non viene specificata che città, o meglio, che maschere di rete escludere della ricerca, ma solo quelle che vogliamo siano considerate. Viene anche specificato anche un altro parametro x di TOPx che rappresenta il numero di destinazione del ranking di scambio di pacchetti.

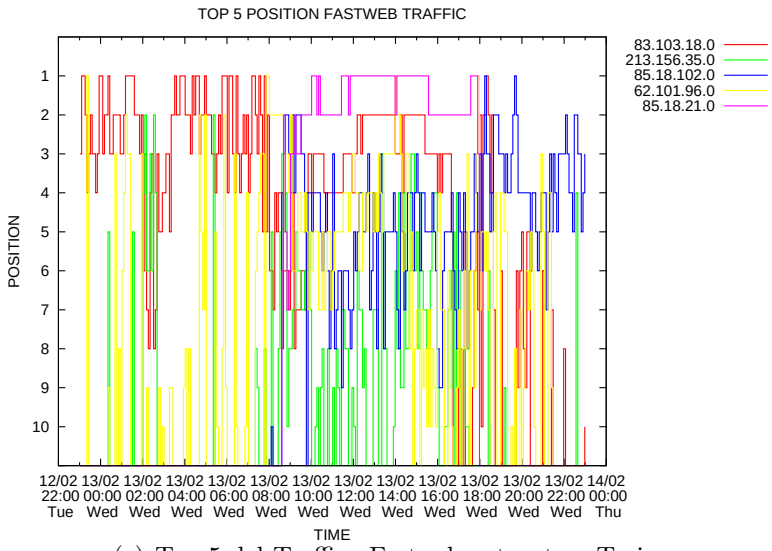
4.4.3 Come ottenere il grafico degli indirizzi IP più visitati

La seguente coppia di comandi ci forniscono i due grafici con il ranking degli IP secondo il parametro configurato all'interno del codice di *matrix.awk*.

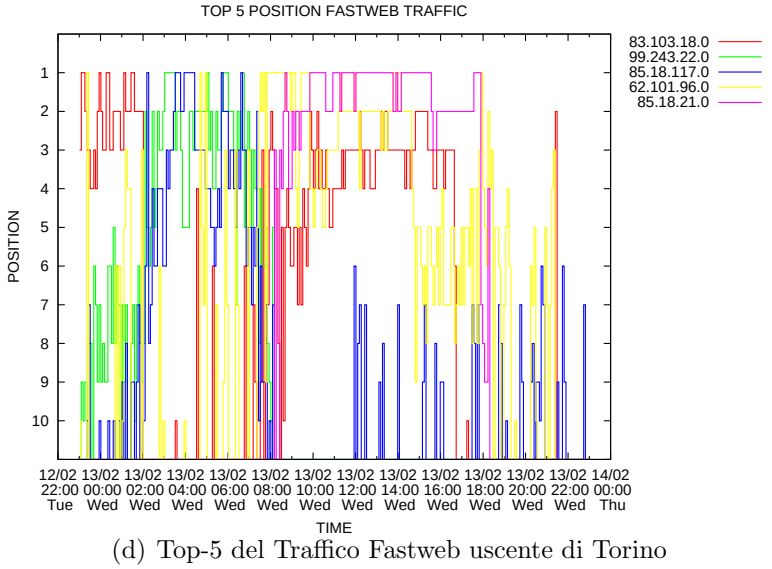
$ ./matrix.awk TOP_IN

$ ./matrix.awk TOP_OUT

Nei seguenti grafici sono illustrati i risultati ottenuti considerando il traffico d'ingresso e di uscita da Torino per 5 indirizzi IP (si è considerato il traffico di mercoledì):



(c) Top-5 del Traffico Fastweb entrante a Torino



(d) Top-5 del Traffico Fastweb uscente di Torino

Capitolo 5

Risultati

5.1 Introduzione

In questo capitolo vengono presentati i graffici dei risultati ottenuti dell'analisi del traffico Internet. Nel primo paragrafo la distribuzione generale del traffico secondo la sua provenenza ad alto livello: il traffico proveniente della rete di Fastweb, traffico di Multicast ed altri tipi di traffico che non provengono della rete di Fastweb.

Nel secondo paragrafo viene analizzata la distribuzione dettagliata per PoP[1]. Ogni PoP corrisponde ad una delle città o nodo importante dell'Italia. Ci sono delle ecezioni, come Milano, che ha più di un PoP dovuto al gran numero di utenti della sua rete. In questo stesso paragrafo vogliamo anche capire se c'è qualche correlazione tra il traffico di un giorno a Torino e gli altri PoP della rete. Perciò si abbiamo calcolato la correlazione utilizzando la retta di regressione e i suoi parametri.

Nel terzo paragrafo è illustrato il traffico organizzato per miniPoPs. Ogni PoP al suo interno è formato per tanti miniPoP. Confronteremo il traffico di tre giorni diversi in diversi periodi di tempo.

Nel quarto paragrafo di questo capitolo è analizzato in dettaglio il traffico di un giorno, cominciando a livello di PoP, poi di miniPoP fino arrivare a livello di sottorete.

Nel quinto paragrafo viene fatto un confronto del traffico durante le ore diurne tra 3 delle principale città italiane. La stessa analisi verrà fatta per il traffico delle ore notturne.

Nel settimo ed ultimo paragrafo di questo capitolo viene elaborato un ranking con i siti più visitati analizzando la sua evoluzione nel tempo.

[1]Point of Presence

5.2 Distribuzione generale del traffico nella rete

Il traffico che si può osservare nella nostra rete viene classificato in tre grandi grupi: il traffico Multicast, il traffico verso nodi della rete di Fastweb ed il traffico che non è destinato ad un utente Fastweb.

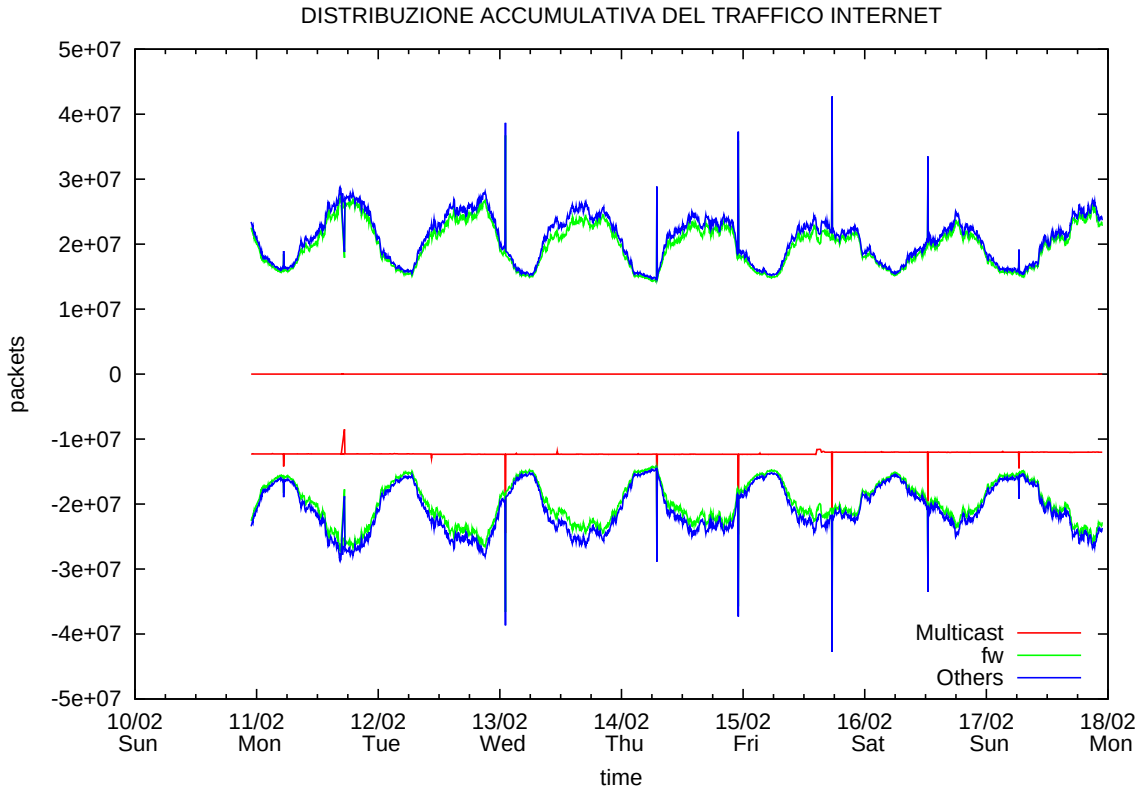


Figura 5.1. Distribuzione assoluta

Nel grafico della Figura 5.1 possiamo vedere questi tre grandi grupi di traffico. Sull'asse negativo delle ordinate è rappresentato il traffico in ingresso al nostro miniPoP, il traffico in uscita invece è sull'asse positivo delle ordinate. Osserviamo che il traffico di Multicast ha solo una componente d'ingresso. Nelle prossime analisi non considereremo questa componente del traffico dato che analizzare il traffico di Multicast non è l'obbietivo di questa tesi .

Nel grafico della Figura 5.2, possiamo vedere la stessa informazione del grafico precendente analizzata in percentuale. Si osserva che il traffico Multicast è una

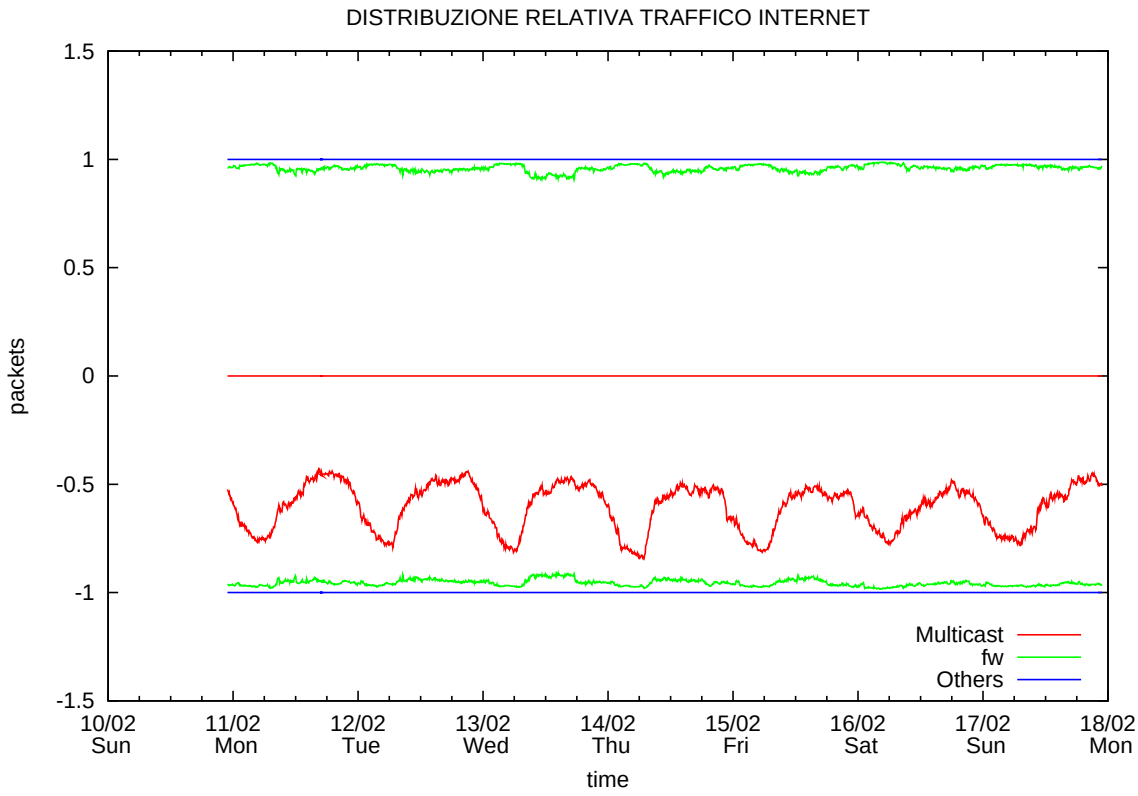


Figura 5.2. Distribuzione relativa

componente importante, tra il 50 ed il 75% del traffico d'ingresso, ed ha un comportamento abbastanza periodico durante la settimana. Se consideriamo anche quello Multicast, il traffico agli altri PoP di Fastweb si riduce al 20-45% del traffico, invece se non consideriamo il traffico Multicast, ad essempio guardando soltanto il traffico che esce del nostro miniPoP, possiamo vedere che rappresenta il 95% del traffico della rete. Resta un 5% circa di traffico verso altri nodi (che non appartengono alla rete Fastweb) che potrebbe scendere sotto il 5% se considerassimo anche il traffico Multicast.

La simetria della rappresentazione del traffico d'ingresso ed uscita è dovuta a due fattori principali. Il primo dovuto al fatto che dobbiamo rappresentare il numero di pacchetti inviati e non la quantità efettiva di bytes. Il secondo è che il protocollo di trasmissione della rete, per ogni paccheto inviato, si deve preocupare di ricevere la conferma di corretta ricezione.

5.3 Distribuzione del traffico per città o PoP

Ogni importante città italiana ha il suo PoP, ci sono delle ecezioni per città come Milano che ne ha più di uno. Ogni PoP è formato da tanti miniPoP al suo interno.

In questa sezione ci centreremo sul traffico che atraversa la nostra rete a livello di PoP. Possiamo vedere l'andamento nel tempo del traffico che entra ed esce del nostro miniPoP verso gli altri PoP della rete di Fastweb.

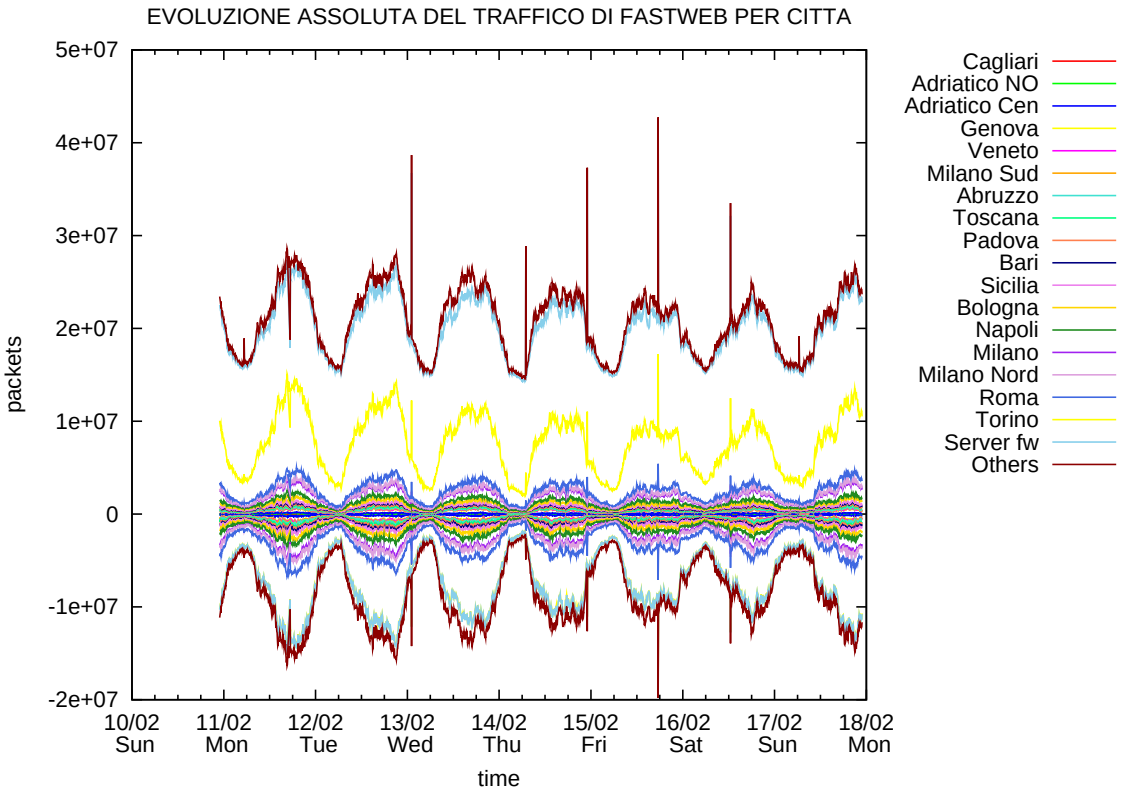


Figura 5.3. Evoluzione assoluta traffico per città

In Figura 5.3 possiamo vedere una asimetria nel traffico d'ingresso rispetto a quello di uscita. La differenza è che nel caso del traffico d'uscita al nostro sniffer c'è il contributo del traffico del Server di Fastweb che si comporta in maniera Multicast, senza un contributo in uscita. Il traffico proveniente dal Server di Fastweb è lo stesso traffico Multicast che abbiamo osservato nelle Figure 5.1 e 5.2. Questo traffico multicast ha due contributi sui nostri grafici: il primo ottenuto considerando un traffico con destinazione multicast, il secondo è ottenuto considerando il server

come origine di traffico iniviato in modo broadcast, allora è ovvio che ne osserviamo il suo contributo anche se è nascosto nel traffico di Fastweb.

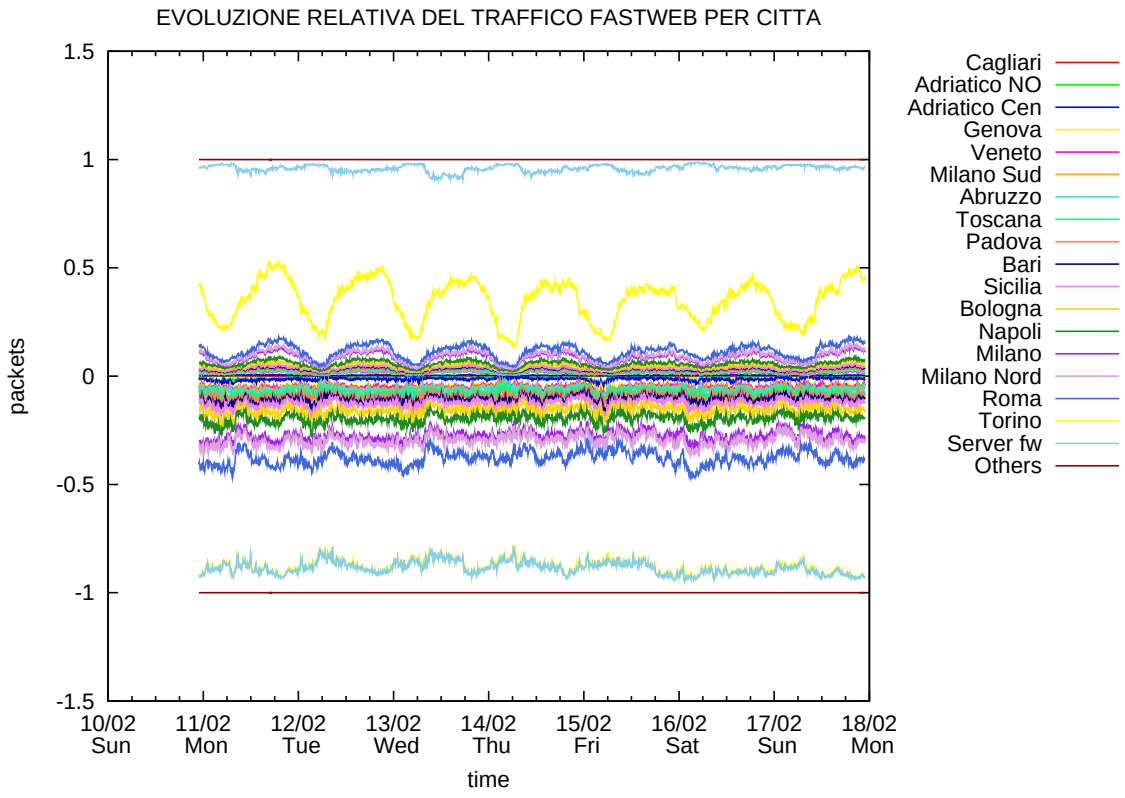


Figura 5.4. Evoluzione relativa del traffico per città

È abbastanza evidente che il traffico del Server di Fastweb è una parte importante del traffico d'ingresso, aprossimatamente intorno al 50-75%, seguito dal traffico destinato a Torino, 15-40%. Il traffico provenienti dagli altri PoP, togliendo quello del Server di Fastweb, e senza considerare il traffico di Torino è intorno al 40%. Invece se consideriamo il traffico proveniente dal Server, il traffico tra PoPs rimane soltanto il 10%.

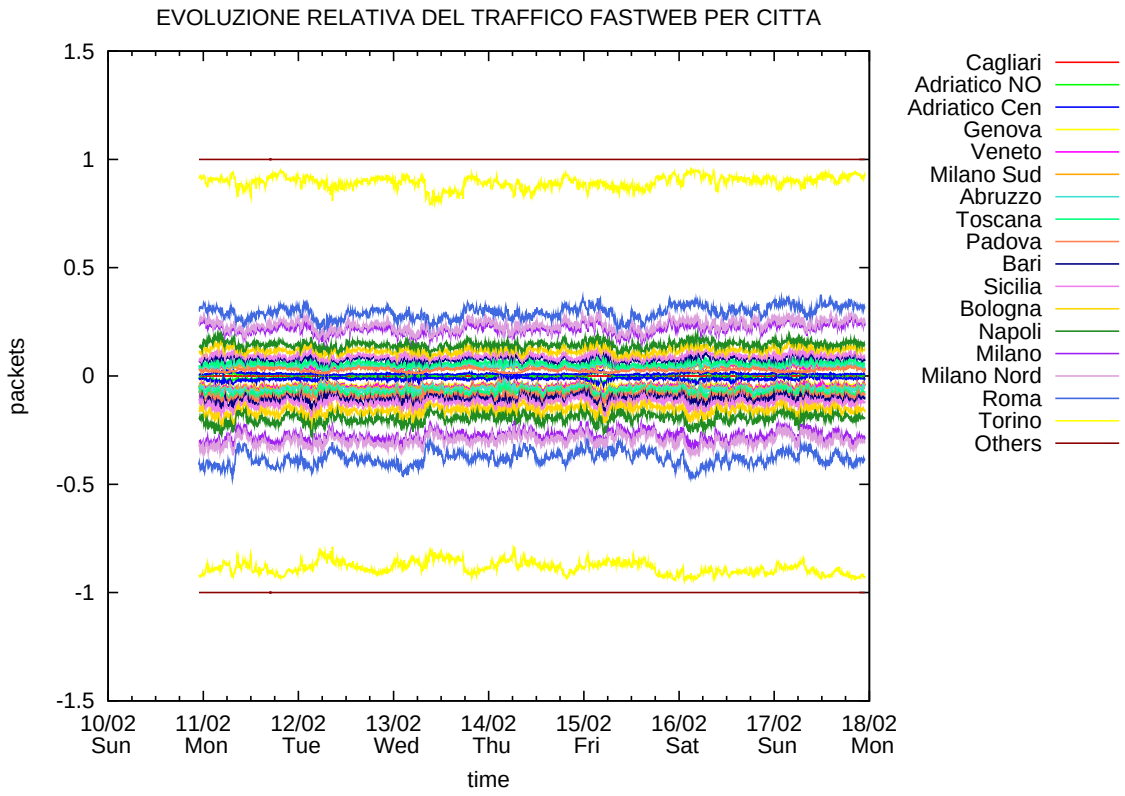


Figura 5.5. Evoluzione assoluta del traffico per città senza il Server di Fastweb

Possiamo osservare la Figura 5.4 che include il traffico del server di Fastweb, e la Figura 5.5 che non contiene il traffico del server.

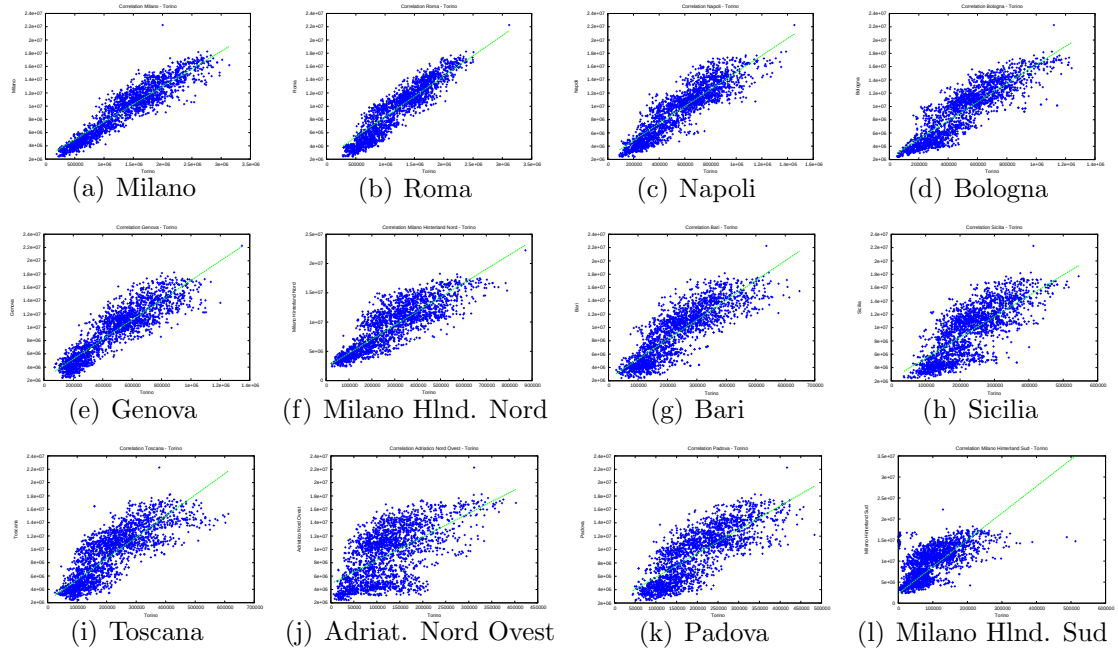
Di seguito potremo vedere la correlazione tra il traffico degli altri PoP col traffico di Torino. La correlazione ci mostrerà quanto si assomigliano i traffici.

5.3.1 Correlazioni fra Torino e gli altri POP

Di seguito possiamo vedere i grafici delle effettuate tra Torino e gli altri PoP della rete di Fastweb. Si può osservare come Milano, Roma, Napoli e Bologna sono delle città che hanno una distribuzione di traffico che assomiglia di più a quella di Torino. Graficamente ce ne possiamo accorgere perché abbiamo poca dispersione dei punti sopra la retta e tutti i punti seguono abbastanza la direzione della retta di regressione.

Grazie ai parametri di regressione della retta nella Figura 5.7, specificamente l'inclinazione m, vediamo come effettivamente Milano è il PoP con m più vicina ad 1. Se l'inclinazione avesse per valore 1 si avrebbe che tanto il traffico di Torino e quello di Milano sono uguali. Per tanto, ci interessano quei valori d'inclinazione il più vicini possibile ad 1.

Ci sono delle città più correlate delle altre, però c'è anche la correlazione con Cagliari che possiamo osservare che è praticamente nulla. Questo significa che il traffico che atraversa il PoP di Torino non assomiglia per niente al traffico che va verso Cagliari. Uno dei motivi di queste differenze può anche venire della distribuzione delle risorse ad ogni città, si potrebbe capire che si ha fatto un investimento maggiore a Milano, ad esempio, che a Cagliari. Anche perché il numero di utenti in entrambe le città è molto diverso.



(a) Milano (b) Roma (c) Napoli (d) Bologna

(e) Genova (f) Milano Hlnd. Nord (g) Bari (h) Sicilia

(i) Toscana (j) Adriat. Nord Ovest (k) Padova (l) Milano Hlnd. Sud

(m) Adriatico centro (n) Abruzzo (o) Veneto (p) Cagliari

Figura 5.6. Correlazioni tra il PoP di Torino e altri PoP della rete di Fastweb

Pop	Final set of parameters	Asymptotic Standard Error
Milano	m = 5.33625 b = 2.2839e+06	+/- 0.04412 (0.8267%) +/- 6.521e+04 (2.855%)
Roma	m = 6.12999 b = 2.28389e+06	+/- 0.06296 (1.027%) +/- 8.072e+04 (3.534%)
Napoli	m = 12.8243 b = 2.2839e+06	+/- 0.1375 (1.072%) +/- 8.388e+04 (3.673%)
Bologna	m =13.9789 b = 2.2839e+06	+/- 0.1649 (1.18%) +/- 9.183e+04 (4.021%)
Genova	m = 14.754 b = 2.2838e+06	+/- 0.1501 (1.018%) +/- 7.959e+04 (3.485%)
Milano Hinterland Nord	m = 23.9536 b = 2.2839e+06	+/- 0.3014 (1.258%) +/- 9.713e+04 (4.253%)
Bari	m = 29.5807 b = 2.2839e+06	+/- 0.3936 (1.331%) +/- 1.028e+05 (4.499%)
Sicilia	m = 31.2412 b = 2.2839e+06	+/- 0.5965 (1.909%) +/- 1.453e+05 (6.362%)
Toscana	m = 31.7494 b = 2.2839e+06	+/- 0.4485 (1.413%) +/- 1.083e+05 (4.74%)
Adriatico Nord Ovest	m = 35.1649 b = 4.84648e+06	+/- 0.9942 (2.827%) +/- 1.416e+05 (2.922%)
Padova	m = 35.5971 b = 2.2839e+06	+/- 0.6084 (1.709%) +/- 1.307e+05 (5.723%)
Milano Hinterland Sud	m = 63.766 b = 2.28387e+06	+/- 1.165 (1.826%) +/- 1.324e+05 (5.799%)
Adriatico Centro	m = 64.7516 b = 4.84648e+06	+/- 1.601 (2.472%) +/- 1.259e+05 (2.598%)
Abruzzo	m = 75.9939 b = 2.2839e+06	+/- 1.205 (1.586%) +/- 1.183e+05 (5.181%)
Veneto	m = 85.0813 b = 2.28384e+06	+/- 1.774 (2.084%) +/- 1.507e+05 (6.599%)
Cagliari	m = 421.473 b = 9.23051e+06	+/- 171.9 (40.77%) +/- 8.498e+04 (0.9206%)

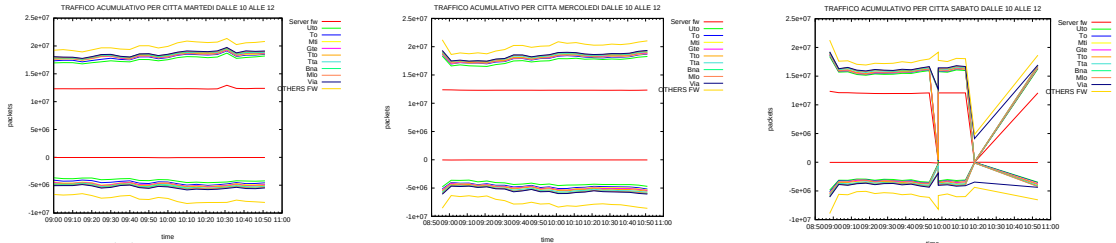
Figura 5.7. Parametri di regressione

5.4 Per mini-Pop

In questa sezione ci interessa guardare il traffico più in dettaglio. È per ciò che analizzeremo tre traffici diversi. Il primo sarà analizzare due ore del mattino di tre giorni diversi

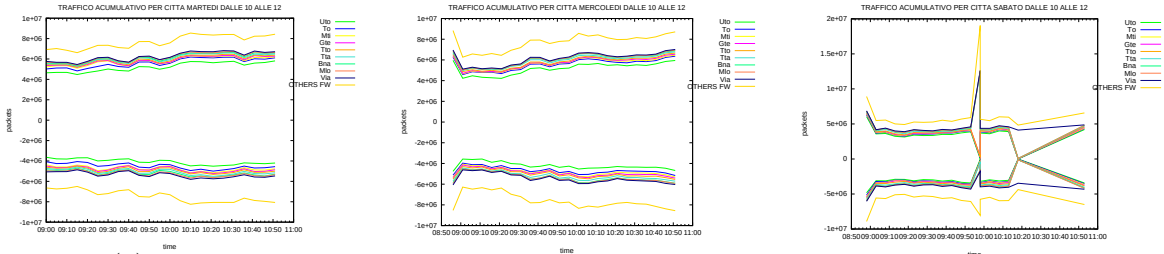
5.4.1 Analisi del traffico di martedì, mercoledì e sabato dalle 9 alle 11

In questi grafici per comodità e una migliore lettura, non abbiamo messo tutti i miniPoP perché ce ne sono tanti, invece abbiamo rappresentato quelli che forniscono una contribuzione maggiore di traffico. Tutti i miniPop che non vengono dettagliati sono compresi nell'insieme di *Others FW*.



(a) Martedì (b) Mercoledì (c) Sabato

Figura 5.8. Traffico per MiniPoP dalle 9 alle 11, con la contribuzione del Server di Fastweb



(a) Martedì (b) Mercoledì (c) Sabato

Figura 5.9. Traffico per MiniPoP dalle 9 alle 11

Come si era già accennato nella sezione precedente possiamo vedere il contributo del traffico del Server di Fastweb soltanto nel traffico in uscita della Figura 5.8. Il traffico che proviene del Server di Fastweb include tutti i servizi che offre Fastweb,

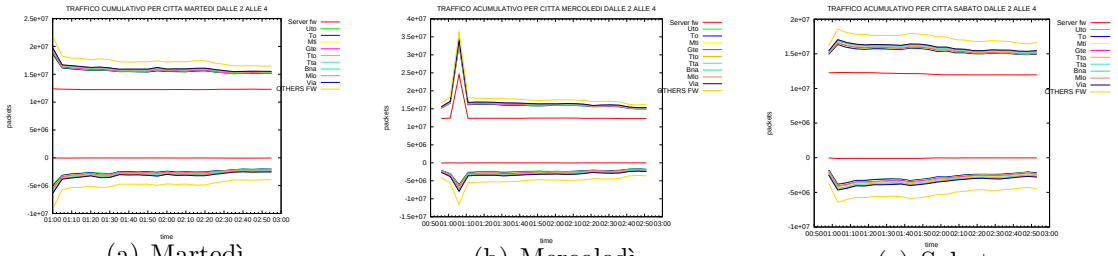
come Video Streamig, VoIP, TV on demand ed altri. Si può vedere come di solito, il traffico proveniente dal Server è pressochè costante, tranne di sabato, quando si presentano due cadute del servizio di Fastweb, probabilmentedovute a malfunzionamenti della rete.

Nella Figura 5.9 abbiamo tolto la componente di Multicast del Server di Fastweb affinchè non ostacoli l'interpretazione dei risultati. Anche cancellando il contributo del Server, si può apprezzare una caduta di alcuni dei miniPoP che offrivano più traffico del sabato, mentre tutti gli altri miniPoP sono sempre abbastanza costanti. Questi picchi sicuramente sono legati alla caduta del Server, causata da un mal funzionamento della rete in quell'istante determinato, e in una regione, giacchè in altre non si sono esperimentate tali problemi.

Un'ultima osservazione da fare è che il traffico nei giorni di martedì e mercoledì il livello di traffico è più elevato se paragonato al traffico del sabato. Cosa normale se consideriamo che durante il fine settimana generalmente non si lavora e si ha, pertanto una riduzione del traffico in rete.

5.4.2 Analisi del traffico di martedì, mercoledì e sabato dalla 1 alle 3

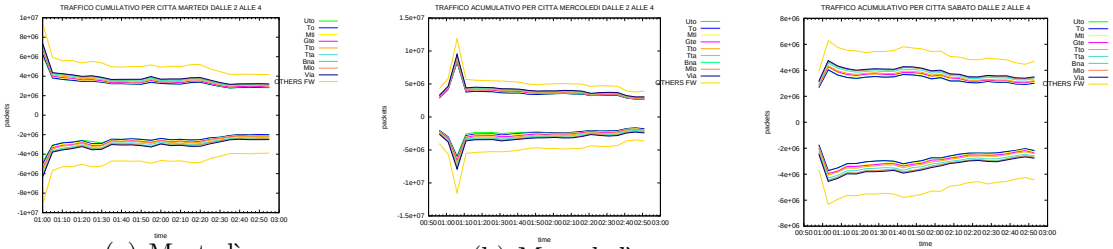
Durante la notte, il traffico è abbastanza costante, se si esclude un leggero calo fino all'una, dopo di ché rimane costante. Un picco inaspettato si verifica mercoledì all'una della notte e lo si osserva sia che si consideri il traffico proveniente dal server di Fastweb o meno. Tale picco, ad una prima osservazione, potrebbe essere dovuto ad una richiesta di un video ad alta definizione.



(a) Martedì (b) Mercoledì (c) Sabato

Figura 5.10. Traffico per MiniPoP dall' 1 alle 3

Anche in questo caso, durante il fine settimana il traffico è più basso che durante la settimana, anche se si esperimenta una diminuzione del traffico progressiva dall'una in poi.



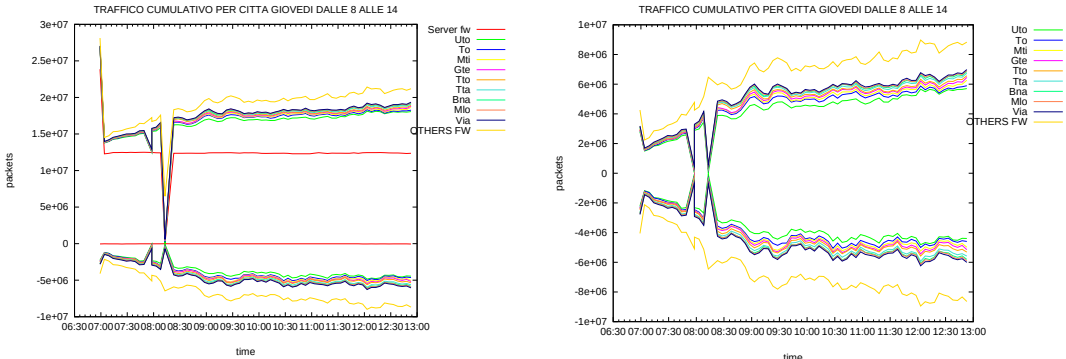
(a) Martedì (b) Mercoledì (c) Sabato

Figura 5.11. Traffico per MiniPoP dall' 1 alle 3

5.4.3 Analisi del traffico di giovedì dalle 7 alle 13

A questo punto non ci resta che analizzare il traffico d'un mattino qualsiasi, ad esempio di giovedì. Osserviamo due cadute della rete: nella prima il server di Fastweb non avverte l'interruzione, nella seconda il server se ne accorge. In entrambi casi quasi tutti i miniPoP sono caduti tranne qualcuno che non si specifica nei grafici che segue ad apportare traffico. Questo miniPoP è contenuto nell'insieme di *Others FW* che come avevamo accennato prima non erano tra i primi contributori de traffico nella rete nell'istante di tempo analizzato.

Se si eccettano le due cadute di servizio della rete, si nota un incremento progressivo del traffico a partire delle prime ore del mattino, fino a tarde mattinata. L'andamento è lo stesso, sia che si consideri il traffico del server (Figura 5.12(a)), sia che lo si scluda (Figura 5.12(b)).



(a) Traffico MiniPop con FW

(b) Traffico MiniPop senza FW

Figura 5.12. Giovedì dalle 7 alle 13

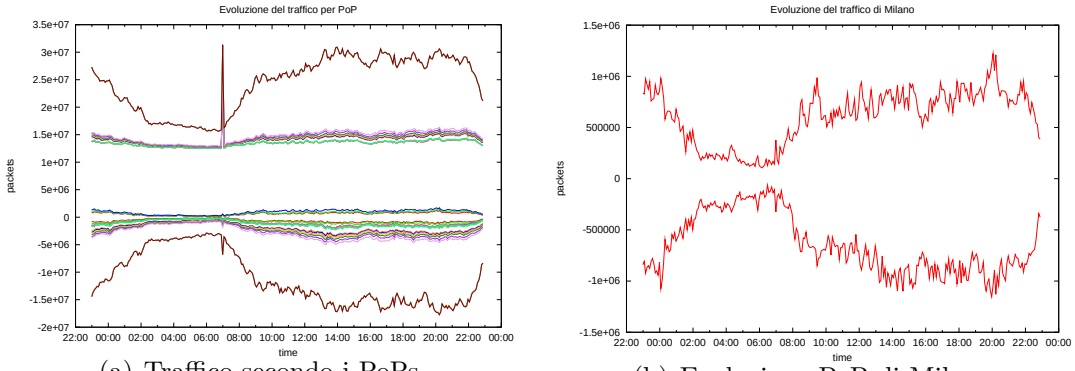
5.5 Analisi del traffico a 3 livelli, PoP, miniPoP, sottorete

L'intento dei paragrafi seguenti è ora quello di analizzare una rete in un intervallo di 24 ore in un giorno a caso, ad esempio di venerdí.

5.5.1 Analisi del traffico verso Milano durante 24 ore

Nel primo grafico 5.13(a) [2] vediamo la distribuzione generale del traffico di un giorno filtrato per PoPs. Di seguito, scegliamo una città, in questo caso Milano, Figura 5.13(b). Il traffico di Milano è alto durante il giorno ed abbastanza basso durante la notte quasi azzerandosi dalle 3 alle 8 circa.

Questo traffico, però, pur essendo abbastanza costante, ha delle fluttuazioni istantanee durante le ore diurne mantenendosi però sopra la media giornaliera, mentre è abbastana costante nelle ore notturne. Ció è dovuto al fatto che il PoP di Milano è formato al suo interno da tanti miniPoP, come si vede nella figura 5.14(a).

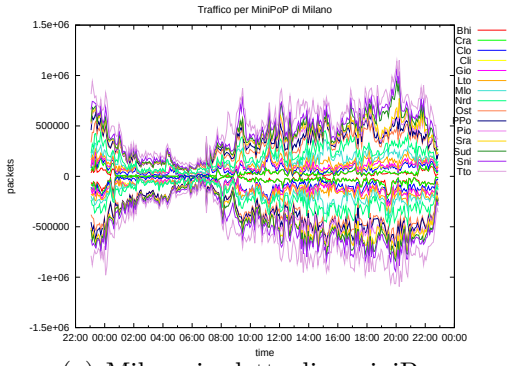


(a) Traffico secondo i PoPs

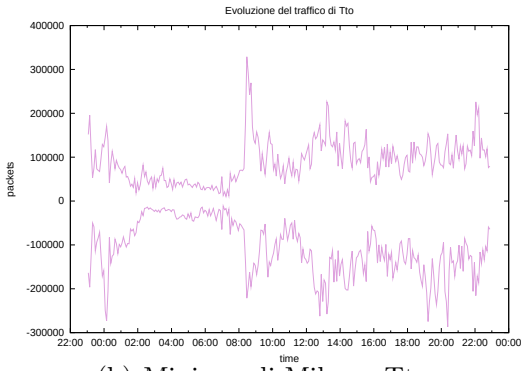
(b) Evoluzione PoP di Milano

Figura 5.13. Traffico secondo i PoPs

[2] per vedere la leggenda ci possiamo rivolgere alla Figura 5.3 in una delle sezioni anteriori.



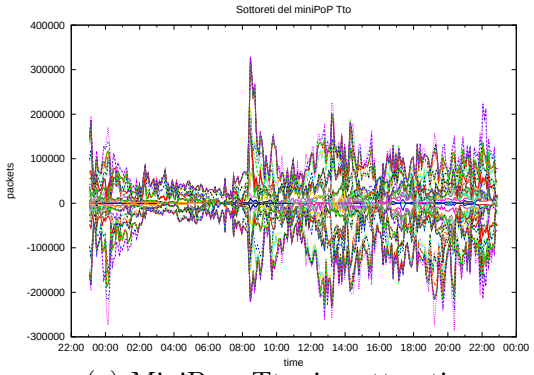
(a) Milano in dettaglio, miniPops



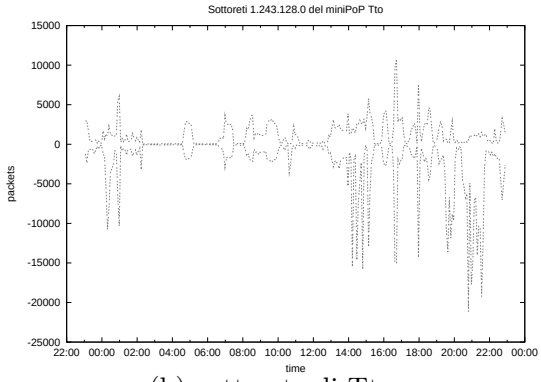
(b) Minipop di Milano, Tto

Figura 5.14. Milano secondo i miniPoPs

Analizziamo un solo miniPoP del PoP di Milano (Figura 5.14(b)). In questo caso, il traffico è più instabile, con oscillazioni più ripide e di diversa intensità, rispetto a quelle ottenute dalla somma di tutti i miniPoP. Nonostante il traffico sia meno costante, come appena osservato, rimane la costante di un carico minore della rete durante la notte.



(a) MiniPop, Tto, in sottoreti



(b) sottorete di Tto

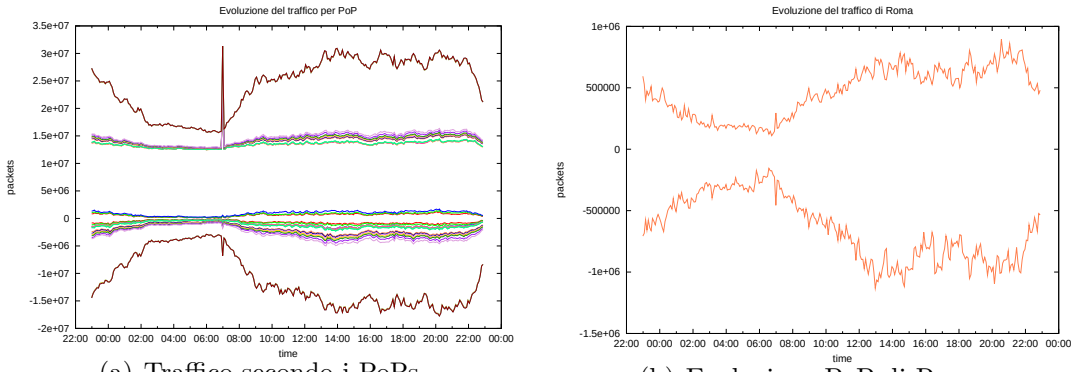
Figura 5.15. Traffico secondo le sottoreti

Ma scendiamo ancora piú in dettaglio. Il miniPoP di Milano, Tto, al suo interno è formato da tantissime sottoreti (Figura 5.15(a)).

Prendendo una sola di queste sottoreti (Figura5.15(b)), vediamo un livello meno costante di traffico, che addirittura si azzera durante la notte mentre subisce forti impennate durante il giorno.

5.5.2 Analisi del traffico verso Roma durante 24 ore

Per quanto riguarda al traffico di Roma, vediamo un comportamento simile a quello di Milano, però con una differenza fondamentale: il traffico di Milano è superiore a quello registrato per Roma.



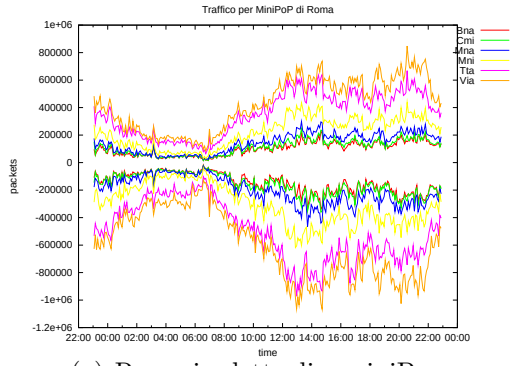
(a) Traffico secondo i PoPs

(b) Evoluzione PoP di Roma

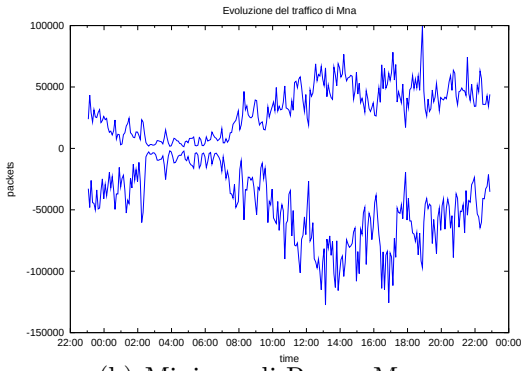
Figura 5.16. Traffico secondo i PoPs di Roma

Un'altra differenza che possiamo apprezzare è la morbidezza della curva durante il passaggio dal giorno alla notte. Nel caso di Milano, la transizione è completata in un intervallo temporale di un'ora, mentre nel caso di Roma, essa avviene in due, tre ore. Si osservano cambiamenti repentini più intensi nel traffico di Milano pertanto.

Guardiamo adesso la Figura 5.17: Mna, risulta essere praticamente una copia del comportamento del PoP di Roma, ma in questo caso abbiamo dei cambiamenti istantani più bruschi e ampiezze un ordine di magnitudine minore. Durante la notte il traffico diventa praticamente inesistente con dei piccoli picchi isolati.



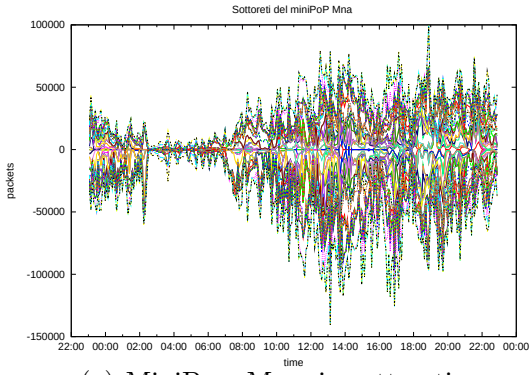
(a) Roma in dettaglio, miniPops



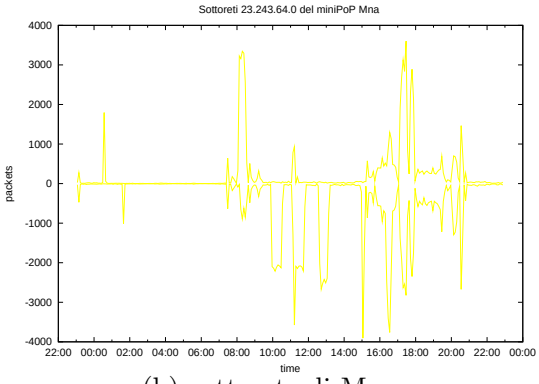
(b) Minipop di Roma, Mna

Figura 5.17. Roma secondo i miniPoPs

Prendiamo adesso una sottorete del miniPoP, Figura 5.18, e vediamo come il suo comportamento è irregolare, con dei picchi altissimi, che quasi raggiungono il massimo del miniPoP, e con dei minimi prossimi a zero. Durante le ore del pomeriggio e della sera possiamo vedere un momento di simmetria di traffico, che non si presenta durante il giorno.



(a) MiniPop, Mna, in sottoreti



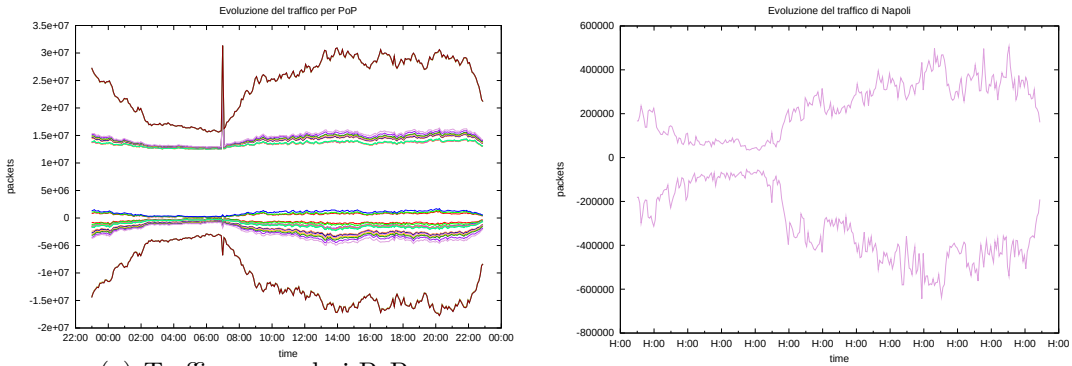
(b) sottorete di Mna

Figura 5.18. Roma secondo le sottoreti

5.5.3 Analisi del traffico verso Napoli durante 24 ore

Il traffico di Napoli è leggermente minore di quello di Roma, Figura 5.19, nottare come il traffico scompare subito dopo mezzanotte, a differenza degli altri PoP dove il traffico non diminuiva fino più entrata la notte.

Tra mezzanotte e le otto del mattino il traffico è quasi inesistente. La transizione giorno-notte è più lenta di Milano e Roma e accade inoltre più tardi che nelle altre metropoli. Durante il giorno si rileva un traffico più o meno costante, se si fa eccezione per alcuni cambiamenti istantanei, tenendosi sempre sulla media dei pacchetti diurni.

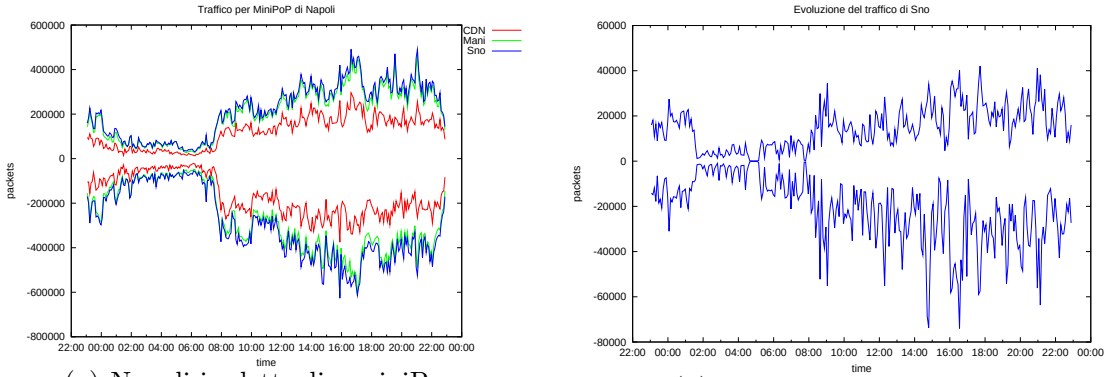


(a) Traffico secondo i PoPs

(b) Evoluzione PoP di Napoli

Figura 5.19. Napoli secondo i PoPs

Per quanto riguarda il traffico diviso per miniPoP, Figura 5.20, si osservano picchi bruschissimi durante le ore diurne, e talvolta nella notte, però d'intensità moderata giaché il traffico notturno è solitamente basso. Durante il giorno ci sono dei cambiamenti istantani costanti.

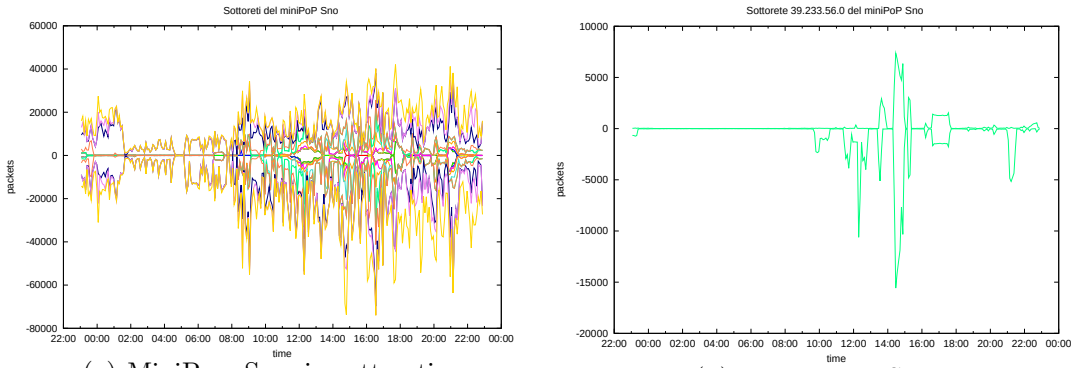


(a) Napoli in dettaglio, miniPops

(b) Minipop di Napoli,Sno

Figura 5.20. Napoli secondo i miniPoPs

Guardiamo adesso la Figura 5.21: possiamo vedere attività internet soltanto tra le ore 10 e 16, durante le altre ore del giorno il traffico di questa sottorete non esiste. Alle 22 notiamo un traffico puntuale soltanto in uscita.



(a) MiniPop, Sno, in sottoreti

(b) sottorete di Sno

Figura 5.21. Napoli secondo le sottoreti

In conclusione, quello che si voleva illustrare in questa sezione è come cambia la distribuzione del traffico quando entriamo più in dettaglio nella rete. Abbiamo visto che il traffico ricevuto ed inviato da un PoP è più o meno costante, tranne le ore della sera che ovviamente diminuisce considerabilmente giachè si tratta di un periodo tradizionalmente di riposo.

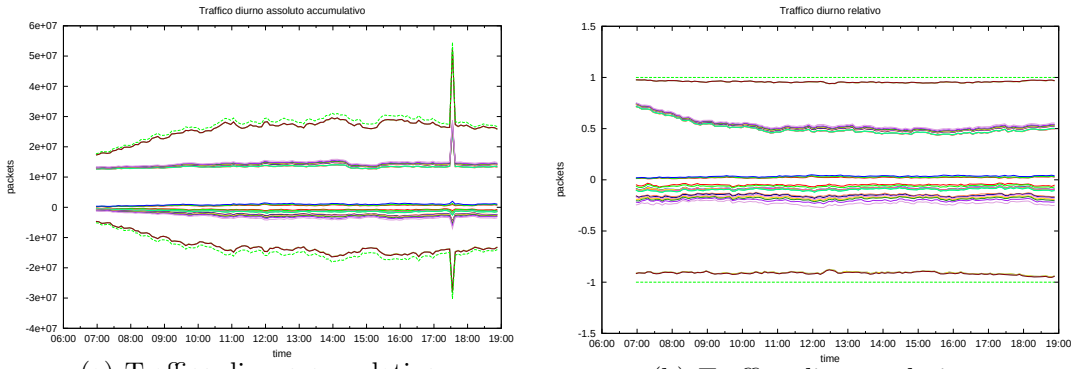
Mano a mano che si entra in dettaglio nel miniPoP il traffico diventa meno costante fino ad arrivare al punto di analisi delle sottoreti che possiamo trovare dei momenti dove scompare il traffico ed altri momenti di traffico intenso in momenti puntuali del giorno.

Come conclusione, diciamo che quanto più in dettaglio siamo, meno prediciibile diventa il traffico.

5.6 Correlazioni giorno-notte secondo i PoP

In questa sezione si vuole illustrare come il traffico giorno-notte tra le diverse città che hanno un flusso più imponente di traffico Internet si assomigliano. Analizeremo come si comportano di giorno e di notte i PoP di Milano, Roma e Napoli rispetto al traffico di Torino. Come giorno di riferimento abbiamo preso un giorno lavorativo, il venerdì.

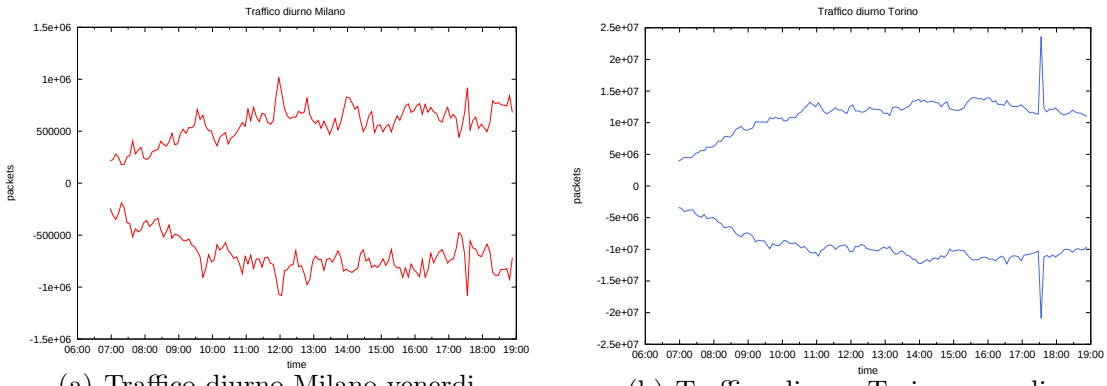
Vediamo prima nel grafico 5.22 la evoluzione assoluta e relativa del traffico per PoP durante le 7 del mattino e le 19 del pomeriggio. Il traffico è abbastanza costante, con un picco importante alle 17:30 del pomeriggio.



(a) Traffico diurno cumulativo

(b) Traffico diurno relativo

Figura 5.22. Traffico diurno totale venerdi

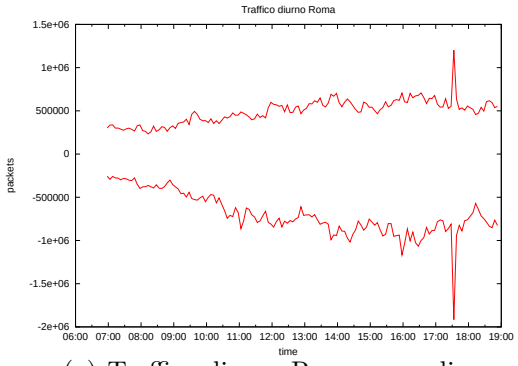


(a) Traffico diurno Milano venerdi

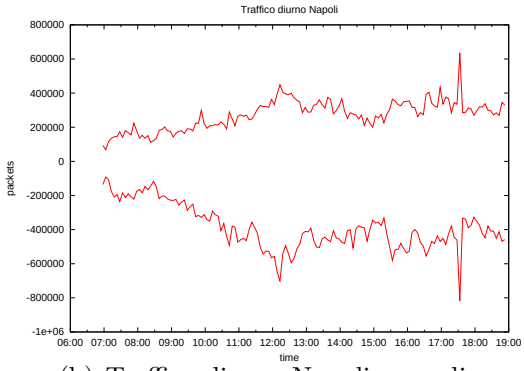
(b) Traffico diurno Torino venerdi

Figura 5.23. Venerdi dalle 8 alle 20

Nella pagina precedente si può osservare il traffico isolato delle città di cui parlavamo prima, nella Figura 5.23 e anche nella Figura 5.24, in questa stessa pagina. Il picco osservato precedentemente nella figura 5.22 è anche osservabile nel traffico di Torino, Roma e Napoli, con l'eccezione di Milano nella stessa ora. D'altra parte, il traffico diurno sembra abbastanza costante in tutti i casi, con l'eccezione anche in questo caso, di Milano che sembra avere un traffico più instabile.



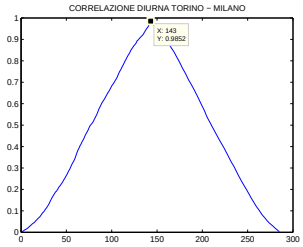
(a) Traffico diurno Roma venerdi



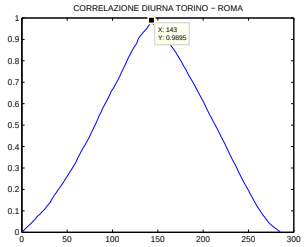
(b) Traffico diurno Napoli venerdi

Figura 5.24. Venerdi dalle 8 alle 20

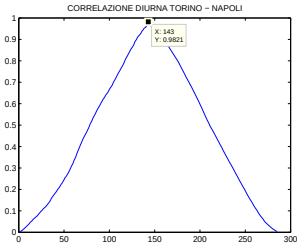
In questo prossimo grafico, Figura5.25, si illustrano le correlazioni del traffico diurno tra le nostre tre città. Ci rivolgeremo soltanto nel massimo di questa correlazione, già che è il momento in cui le due città saranno alineate l'una sopra di l'altra mostrandoci in tanto per uno quanto assomigliano le diverse forme del traffico.



(a) Torino - Milano



(b) Torino - Roma



(c) Torino - Napoli

Figura 5.25. Correlazioni diurne di Torino con gli altri PoP più importanti

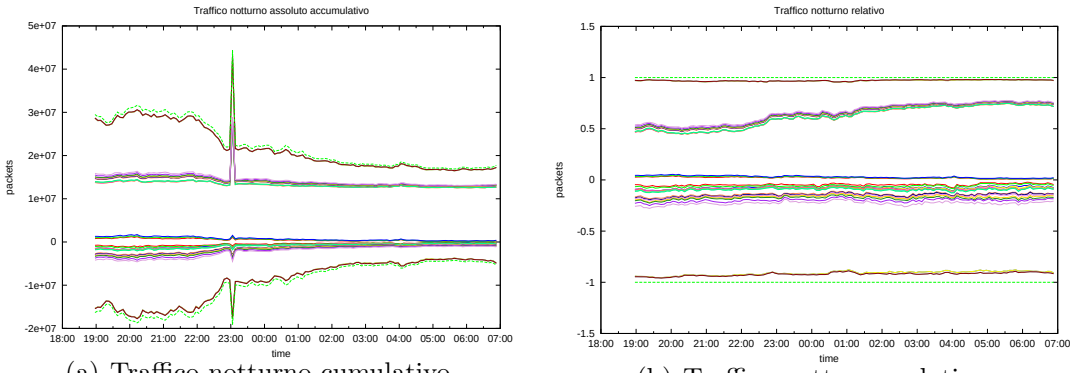
Nella prossima tavola possiamo vedere un riassunto più chiaro di questi massimi di correlazione:

Pop	Massimo correlazione
Milano	0.9852
Roma	0.9895
Napoli	0.9821

Figura 5.26. Correlazioni diurne

Si vede che il traffico per quanto riguarda la forma assomiglia in tutti e tre casi in più del 98%. Questo è un risultato normale, anche se sembra stranno che possano assomigliare tanto, però dobbiamo avere in mente che stiamo analizzando la forma del traffico e non la quantità di dati. Un'altra cosa da tenere presente è che queste correlazioni sono normalizate all'energia di entrambi i segnali.

A continuazione si mostra il grafico di traffico notturno:



(a) Traffico notturno cumulativo

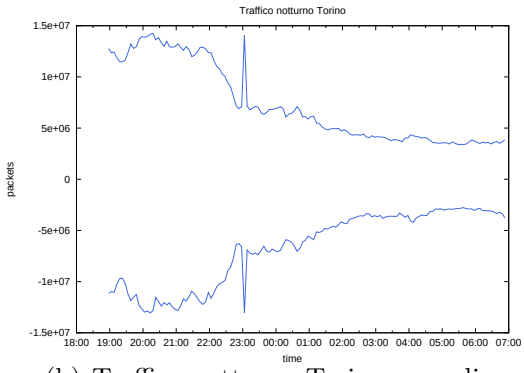
(b) Traffico notturno relativo

Figura 5.27. Traffico notturno totale venerdì

In questo caso si apprezza che il traffico non è così costante come abbiamo visto durante le ore diurne. Il traffico si mantiene costante fino alle 22 quando comincia a scendere drasticamente tra le 22 e le 23. Si osserva un picco importante e puntuale alle 23 e poi continua a scendere raggiungendo un traffico costante dopo le 2 della notte.

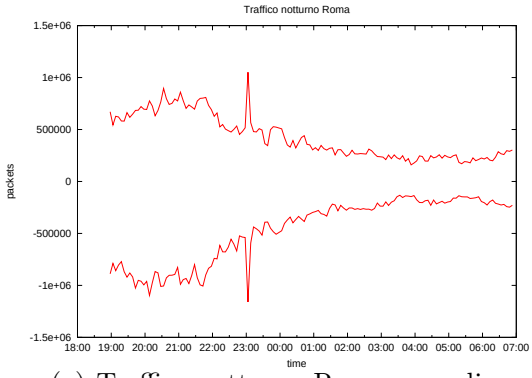
Vediamo adesso di maniera dettagliata ogni città isolata dalle altre nella Figure 5.28 e la Figura 5.29:

(a) Traffico notturno Milano venerdi

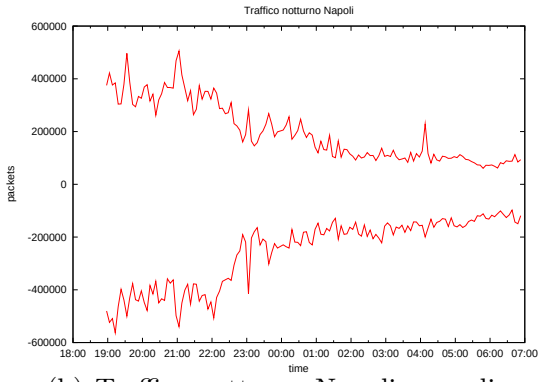


(b) Traffico notturno Torino venerdi

Figura 5.28. Venerdi dalle 20 alle 8



(a) Traffico notturno Roma venerdi

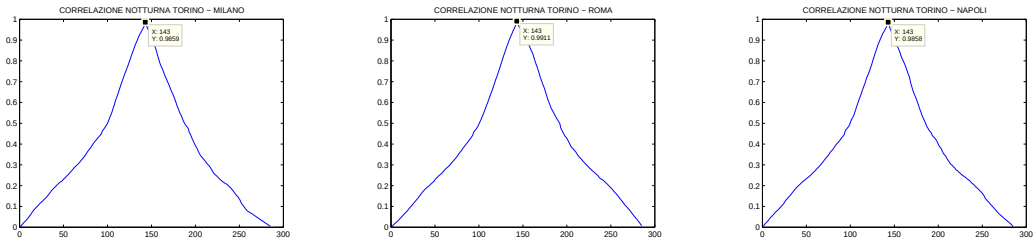


(b) Traffico notturno Napoli venerdi

Figura 5.29. Venerdi dalle 20 alle 8

Anche se isolate, possiamo vedere questo picco delle 23 in ogni città, anche se quello di Napoli è abbastanza nascosto e si intuisce soltanto nel traffico di uscita. Entrambe le città hanno un comportamento simile con una discesa progressiva raggiungendo dei valori minimi a notte addentrata.

Di seguito, i grafici della correlazione evidenziano quanto il traffico diurno è simile a quello notturno:



(a) Torino - Milano (b) Torino - Roma (c) Torino - Napoli

Figura 5.30. Correlazioni notturne Torino con gli altri PoP importanti

Pop	Massimo correlazione
Milano	0.9859
Roma	0.9911
Napoli	0.9858

Figura 5.31. Correlazioni notte

La correlazione indica che il traffico più somigliante a quello di Torino è quello proveninente da Roma, avvicinandoci a percentuali del 99% anche se per gli altri due adsi, raggiungiamo percentuali del 98%.

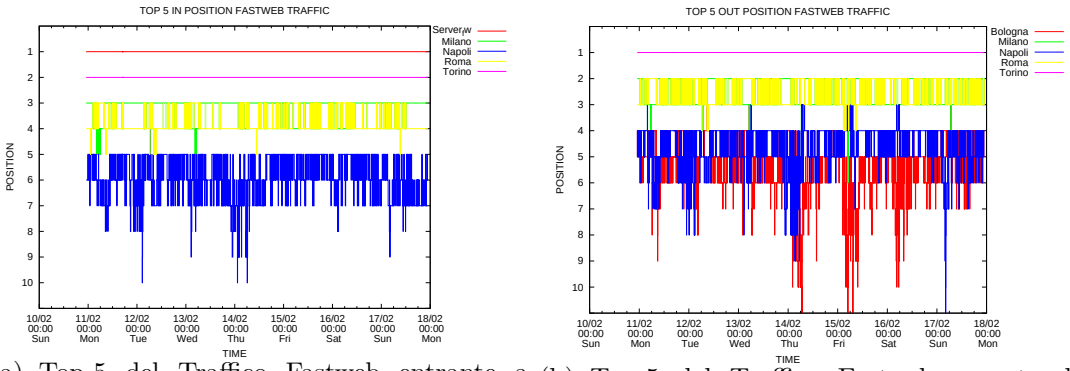
La conclusione che si trae da questa sezione è che le tre città che hanno uno importante scambio di traffico con Torino, assomiglia sia per forma sia per quantità di pacchetti scambiati. Le correlazioni sono altissime sia di girono sia di notte.

5.7 Risultati dei siti più visitati

In questa sezione possiamo mettere in evidenza l'evoluzione dei siti più visitati, la constanza nel tempo ed il cambiamento dipendente dal giorno della settimana preso in esame.

5.7.1 TOP5 Traffico Fastweb

Prima di tutto diamo uno sguardo generale al traffico Fastweb a livello di PoP. Vediamo con chi è che il PoP di Torino scambia più traffico.



(a) Top-5 del Traffico Fastweb entrante a Torino

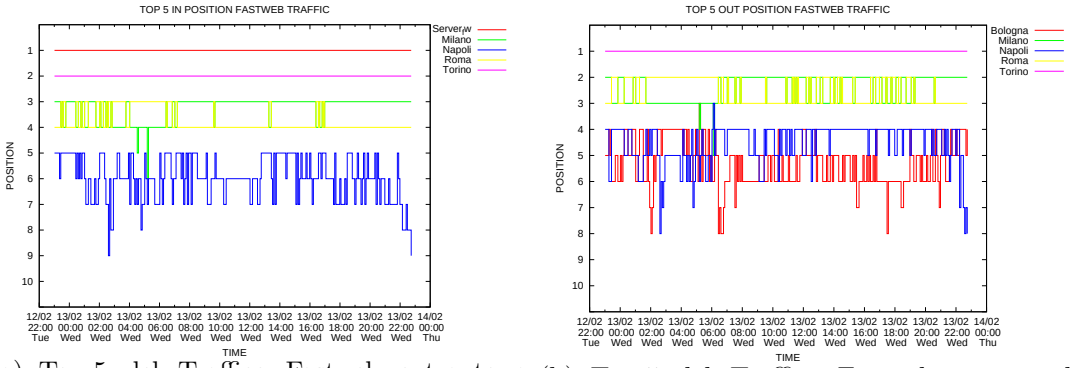
(b) Top-5 del Traffico Fastweb uscente da Torino

Figura 5.32. Top 5 del Traffico Fastweb IN e OUT di Torino per tutta la settimana

Si può osservare che lo scambio a livello di PoP risulta abbastanza costante e che risulta essere costante anche il traffico in ingresso. Il PoP di cui riceviamo più traffico è quello del server di Fastweb e tale risultato è coerente con i risultati ottenuti fino adesso, che evidenziavano questa quantità enorme di traffico proveniente dal server di Fastweb. Da notare che nel traffico d'uscita il Server di Fastweb non appare e che risulterebbe sicuramente al primo posto se non avessimo tolto la componente di Multicast.

Dopo il Server di Fastweb, il PoP di Torino risulta essere il secondo per flusso di traffico entrante ed uscente a Torino. Il terzo ed il quarto posto sono occupati dal PoP di Roma e Milano, anche se non è facilmente distinguibile chi occupi il terzo e chi il quarto posto. Come novità, per quanto riguarda il traffico in uscita abbiamo Bologna nel quinto posto dei PoP più visitati. Per trovare riscontro possiamo guardare i grafici della Figura 5.32.

Verifichiamo ora se il traffico, i siti più visitati, si mantengono costanti se analizziamo un solo giorno della settimana.

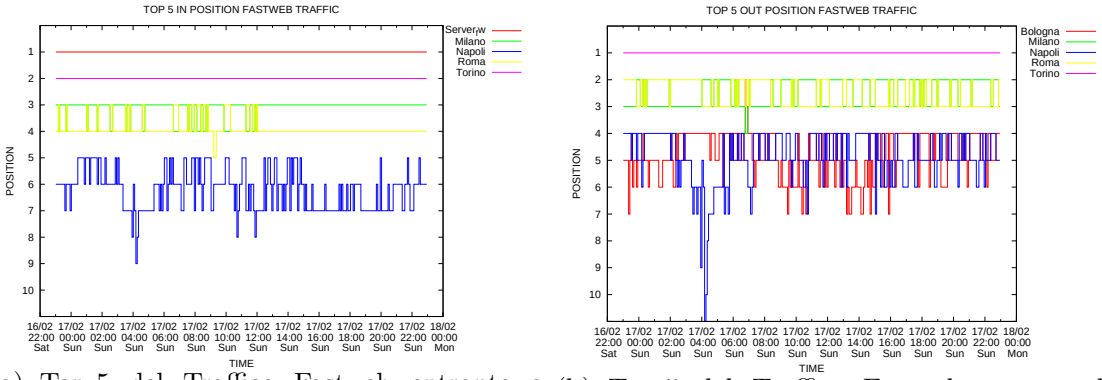


(a) Top-5 del Traffico Fastweb entrante a Torino

(b) Top-5 del Traffico Fastweb uscente da Torino

Figura 5.33. Top 5 del Traffico Fastweb IN e OUT Torino Mercoledì

Si può notare che Milano, nel giorno di mercoledí, presenta un traffico d'ingresso più elevato di quello di Roma. Non ci sono particolari differenze per il traffico in uscita anche se si evidenzia una lieve predominanza da parte del PoP di Roma. Napoli e Bologna rispettivamente occupano il quarto ed il quinto posto.



(a) Top-5 del Traffico Fastweb entrante a Torino

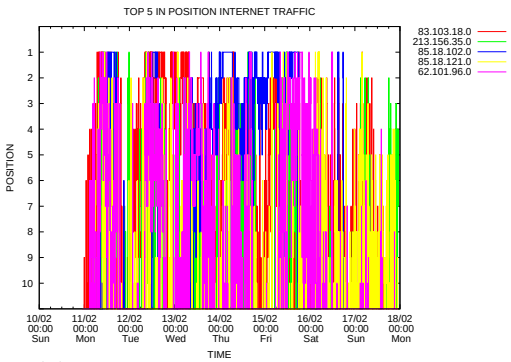
(b) Top-5 del Traffico Fastweb uscente da Torino

Figura 5.34. Top 5 del Traffico Fastweb IN e OUT Torino Domenica

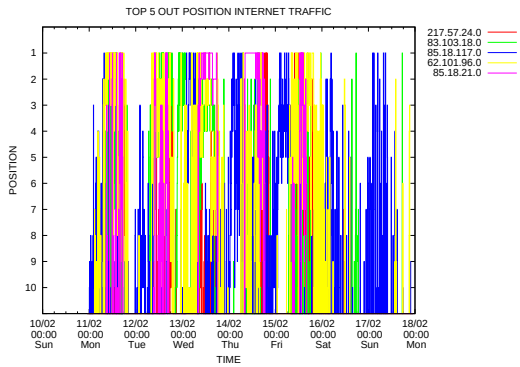
Nella giornata di domenica, il traffico in uscita di Bologna e Napoli risulta meno elevato. Torino occupa il primo postao mentre Roma e Milano, seppure con lievi differenze, il secondo ed il terzo.

5.7.2 TOP5 Traffico Internet

Dopo aver esaminato il traffico a livello di PoP passiamo ad analizzare più in approfonditamente le sottoreti. In questa sezione verrà evidenziato qualli sottoreti sono quelle più visitate durante la settimana, sarà possibile notare come adesso esistono più cambiamenti con rispeto al traffico analizzato a livello di PoP. Come dimostrato in precedenza, quanto più in dettaglio si entra, meno predicibili diventano i siti visitati.



(a) Top-5 Traffico Internet IN Torino



(b) Top-5 Traffico Internet OUT Torino

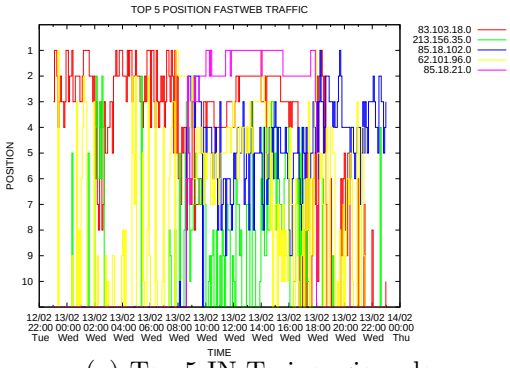
Figura 5.35. Top 5 del Traffico Internet IN e OUT di Torino per tutta la settimana

Dal grafico di tutta la settimana, dobbiamo estrarre gli indirizzi IP più visitati degli utenti. Queste misure sono state fate prendendo come tempo di campione tutta la settimana, perciò se analizzasemo qualle sono gli indirizzi IP del TOP 5 del mercoledì, pero ponderando soltanto le probabilità del mercoledì, ottendremo lo stesso risultato che quello ottenuto per tutta la settimana? Cambierà se invece di essere mercoledì il giorno scelto, scegliamo la domenica? A queste domande verrà data risposta di seguito.

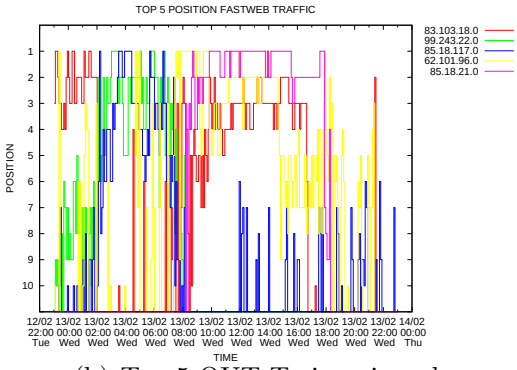
Durante la settimana abbiamo visto che i siti più visitati sono:

Posizione	IP	sito web	IP	sito web
1	83.103.18.0	FW SMALL BUSINESS	217.57.24.0	WEB SINERGY SOLUTION
2	213.156.35.0	FW POP 4105	83.103.18.0	FW SMALL BUSINESS
3	85.18.102.0	FW POP 4100	85.18.117.0	FW BLUVAL SRL
4	85.18.121.0	RADIO SETTIMO	62.101.96.0	RADIO SETTIMO
5	62.101.96.0	FW POP 4102	85.18.21.0	FW POP 4100

Verifichiamo se veramente questi siti sono i più visitati se prendiamo in analisi soltanto il mercoledí :



(a) Top-5 IN Torino, giornale

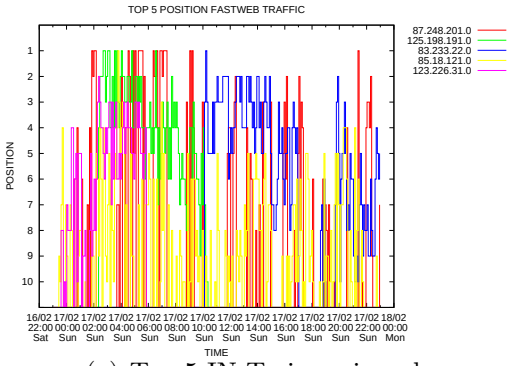


(b) Top-5 OUT Torino giornale

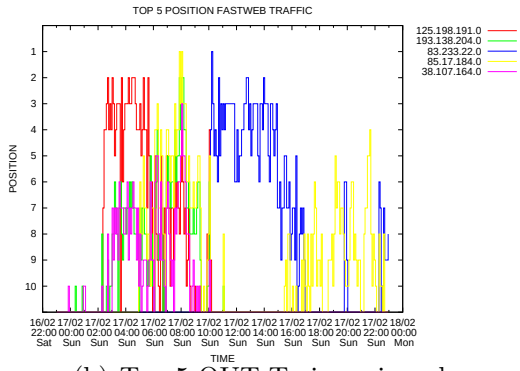
Figura 5.36. Top 5 del Traffico Fastweb IN e OUT Torino Mercoledi

Posizione	IP	sito web	IP	sito web
1	83.103.18.0	FW SMALL BUSINESS	83.103.18.0	FW SMALL BUSINESS
2	213.156.35.0	FW POP 4105	99.243.22.0	CANADA CABLE TLC
3	85.18.102.0	FW POP 4100	85.18.117.0	FW BLUVAL SRL
4	62.101.96.0	FW POP 4102	62.101.96.0	RADIO SETTIMO
5	85.18.21.0	FW POP 4100	85.18.21.0	FW POP 4100

Possiamo nottare che mercoledì il ranking generale dei siti visitati corrisponde anche, con delle piccole variazioni, al traffico calcolato soltanto mediando i dati di un giorno. Una cosa importante da osservare è che anche se tutto il traffico era stato filtrato il traffico che era destinato a Fastweb, continua ad essere dominio di Fastweb. Sono degli indirizzi IP che appartengono a Fastweb però che sono state affittate per parte di altre aziende.



(a) Top-5 IN Torino, giornale



(b) Top-5 OUT Torino, giornale

Figura 5.37. Top 5 del Traffico Fastweb IN e OUT Torino Domenica

Pos	IP	sito web	IP	sito web
1	87.241.201.0	NETHER.LIMELIGHT NET	125.198.191.0	TOKIO NEC CORP.
2	125.198.191.0	TOKIO NEC CORP	193.138.204.0	NETHER. KADIS NET.
3	83.233.22.0	STOCKHOLM BROAD.	83.233.22.0	STOCK. BROAD.
4	85.18.121.0	RADIO SETTIMO	85.17.184.0	AMSTER. LEASEWEB
5	123.226.31.0	TOKIO OPEN COMP NET.	38.107.164.0	US PSINet

Se invece guardiamo i siti più visitati nella domenica, possiamo vedere come sono totalmente diversi da quelli calcolati durante tutta la settimana. Adesso non c'è nessuna rete gestita da Fastweb, sono tutte delle aziende internazionali. Probabilmente possono essere dei Servers di video streaming.

Capitolo 6

Conclusioni

Lo scopo di questa tesi era proporre una tecnica per analizzare dei dati estratti da uno sniffer situato all'interno della propria rete Fastweb del Politecnico. Si è proposta una metodologia da seguire nel caso di volere analizzare altri dati ottenuti in una forma simile. La finalità era essere in grado di trovare una riga ed una colonna della matrice di traffico della rete dai dati estratti, dopo un ulteriore analisi ed elaborazione.

Se in un futuro ci fosse uno sniffer in ogni PoP della rete di Fastweb e fosse applicata la stessa metodologia potremmo essere in grado di stimare completamente la matrice di traffico e non solo una riga ed una colonna. Data l'alta complessità di questa proposta se mettiamo insieme i risultati otenuti per il PoP di Torino ed altri studi fati precedentemente, si potrebbero estrapolare i risultati a tutta la rete FW.

Altri lavori da condurre in futuro possono essere orientati a creare una interfaccia grafica per la generazione automatica dei grafici allo stesso modo in cui funziona il sito di Tstat, http://tstat.tlc.polito.it/web.shtml

L'informazione ricavata da queste misure della rete può essere utile in un futuro studio più esteso per adattare le infrastrutture che Fastweb ha sparse per il territorio italiano, in modo da ottimizare le risorse, determinare nuovi o diversi instradamenti, etc. In definitiva, migliorare la qualità del servizio per l'utente finale in cambio di un uso più efficente della rete.

Bibliografia

[1] Leslie Lamport.*LaTex : A document Preparation System.*Addison-Wesley, 1986.

[2] Christian Rolland.*LaTex guide pratique.*Addison-Wesley, 1993.

[3] J. Kowalski and B. Warfield. *Modeling traffic demand between nodes in a telecommunications network*. In ATNAC95, 1995.

[4] Y. Vardi. *Network tomography: estimating source-destination traffic intensities from link data*. Journal of the American Statistical Association, pages 365.377, 1996.

[5] J. Cao, D. Davis, S. V.Wiel, and B. Yu. *Time-varying network tomography.* Journal of the American Statistical Association, 95:1063.1075, 2000.

[6] A. Soule, A. Lakhina, N. Taft, K. Papagiannaki, K. Salamatian, A. Nucci, M. Crovella, and C. Diot. *Traffic matrices: Balancing measurements, inference and modeling*. In SIGMETRICS'05, Banff, Canada, 2005.

[7] S. McCanne, C. Leres and V. Jacobson, *libcap*, http://www.tcpdump.org, 2001.

[8] G. Iannaconne, C. Diot, I. Graham, A.McGregor, and M. Pearson,*Design principles for accurate passive measurement, Proc PAM2000: The first Passive and Active Measurement Workshop, Hamilton, New Zeland*, pp. 1-7, April 2000.